U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

MSc

2.º
CICLO

FCUP
2015

U.PORTO

Planctomycetes attached to algal surfaces: Insight into their genomes

Mafalda Seabra Faria

FC

Planctomycetes attached to algal surfaces: Insight into their genomes

Mafalda Seabra Faria

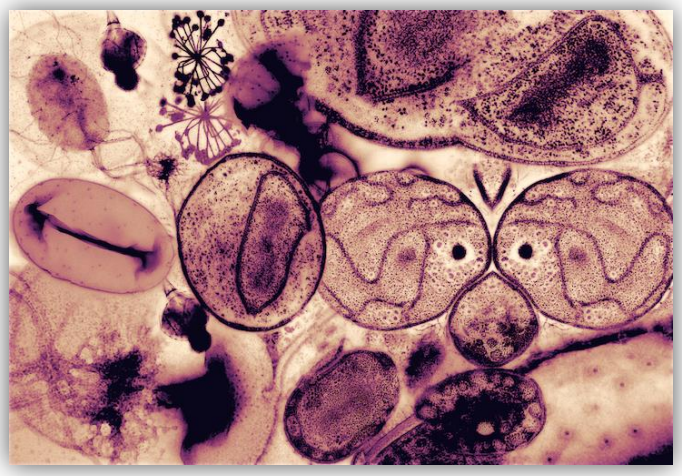
Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto
Laboratório de Ecofisiologia Microbiana da Universidade do Porto

Biologia Celular e Molecular

2014/2015

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc



Planctomycetes attached to algal surfaces: Insight into their genomes

Mafalda Seabra Faria

Mestrado em Biologia Celular e Molecular
Biologia
2015

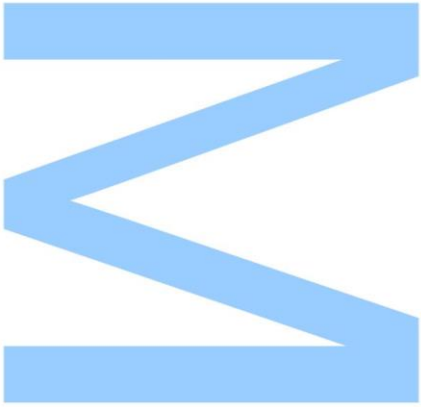
Orientador
Olga Maria Oliveira da Silva Lage, Professora Auxiliar, Faculdade de Ciências da Universidade do Porto

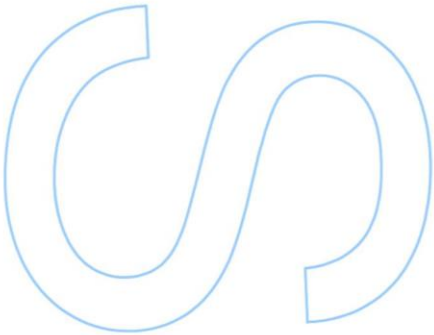
Co-orientador
Jens Harder, Senior Scientist and Professor, Max Planck Institute for Marine Microbiology

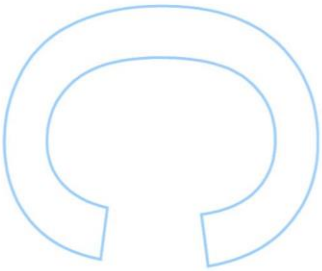
Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____







"Tell me and I forget, teach me and I may remember, involve me and I learn."

Benjamin Franklin

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Professor Olga Maria Lage for the continuous support, enthusiasm, patience and guidance for not only my thesis, but for the past 3 years. Her guidance and motivation were essential for the development of this work. Also, I would like to thank her for all the inspiring and developing opportunities that she was able to offer me, either in Erasmus by getting to know other practices, experience diversity and the unexpected or, by simply being able to attend an International congress, getting to know more and more the scientific world around me.

Also, I would like to thank my entire fellow the lab mates from LEMUP that, over the years, I have had the pleasure to share the lab with. A special thanks to Patricia! Thanks for putting up with all my insecurities, outpourings, disbeliefs, crazy moments and endless happy days as well. Her encouragement, insightful comments help me made all this way easier and way happier!

My sincere thanks also goes to Professor Jens Harder, PhD student Jana Kizina and all the other people I met in the Max Planck Institute in Bremen during the Erasmus internship. Their help, warm welcome and guidance through the three months spent there will always be remembered. *Ich danke euch für alles!*

Dr. Damien Devos and PhD student Nicola Bordin, for the insightful week spent in Sevilla in their Bioinformatics lab. Their endless help and knowledge were of great value for the last developments of this thesis. It was great to feel part of the group, even if for a week. *iMuchas gracias!* .

I couldn't not thank my NGO of these past 5 years, BEST and, specially, Local BEST Group Porto. Thanks for all the extracurricular work, awesome energy, happy moments, (good) friends met and also, for the bunch of challenges I've been lucky to experience. They shaped me into the person who I am today and, for sure, it wouldn't be possible without them. "Work hard, party harder!"…and so we did. To my awesome international team of Regional Advisers 14/15, the x-RAys, this one is for you as well! You were indeed a huge support during the hardest and greatest moments and I am so happy I had every single one of you as part of my team!

To all my dear and good friends, you certainly know how much you're important to me. And for sure, you know who you are! ;) *Biomelgas, BESTies, amigos com mais de duas décadas e turma M:BCM 13/14… agradeço-vos por todos os momentos passados, todos eles contribuíram para o finalizar de mais um objectivo!*

Last but certainly not least, I would like to thank all my family, especially, my mother for the endless care, support and sleepless nights over the past 23 years; my father for the unconditional support and appreciation and my grandparents, for all the pampering, sweet words and for bringing me up. *Muito obrigada por tudo, estas palavras nunca poderão expressar o quão agradecida eu estou a cada um de vocês por todas as oportunidades, apoio e carinho <3*

Resumo

O filo dos *Planctomycetes* é um grupo notável de bactérias com características morfológicas e celulares fascinantes e fora do comum. Este filo é conhecido pela sua relevância nas áreas da biologia celular, evolução e ecologia.

Hoje em dia, com o desenvolvimento de ferramentas bioinformáticas mais rápidas e de fácil uso, na análise de sequenciação de genomas é possível estudar microorganismos analisando directamente os seus genomas, mesmo se não cultivados.

Sequências não tratadas das três estirpes de planctomycetes isolados das superfícies das macroalgas, *Rubripirellula obstinata* estirpe LF1, *Roseimaritima ulvae* estirpe UC8 e uma estripe não caracterizada, a estirpe FC18, foram obtidas após sequenciação com a tecnologia Illumnia MiSeq. Seguidamente, essas sequências foram submetidas ao *software* SPAdes para montagem dos respetivos genomas e anotadas nos programas RAST e Prokka.

Análises compreensivas e comparativas destes três genomas foram realizadas contra os genomas das *Rhodopirellula baltica* SH1^{T}, *Blastopirellula marina* DSM 3645 e *Planctomyces limnophilus* DSM 3776. As características gerais dos genomas das três estirpes mostraram ser concordantes com outros genomas dos *Planctomycetales*. Um estudo para a detecção de genes relacionados com a produção de compostos secundários foi realizada pelo programa antiSMASH e mostrou a presença de genes promissores que transcrevem moléculas putativas com actividade antimicrobiana.

As análises genómicas comparativas das estirpes LF1, FC18 e UC8 mostraram ainda a presença de proteínas únicas que poderão estar relacionadas com o microambiente a partir do qual as estirpes foram isoladas, o complexo biofilme das macroalgas.

Palavras-Chave: Planctomycetes, *Rubripirellula obstinata*, LF1, *Roseimaritima ulvae*, UC8, estirpe FC18, *Rhodopirellula baltica, Blastopirellula marina, Planctomyces limnophilus,* Bioinformatica, Sequenciação, Montagem genómica, Anotação, Genoma, Actividade antimicrobiana, Biofilmes

Abstract

Planctomycetes is a remarkable group of bacteria with unusual and striking cellular and morphological features. They are acknowledged for their meaningful relevance in the fields of cell biology, evolution and ecology.

Nowadays, with the development and support of faster and user-friendly *in silico* tools, used in genome sequencing analysis it is possible to study microorganisms by looking directly at their genomes, even if not cultivated. This enables a deeper insight and understanding into biology of microorganisms.

Raw sequences from three strains of planctomycetes isolated from algal surfaces, *Rubripirellula obstinata* strain LF1, *Roseimaritima ulvae* strain UC8 and a yet to characterize strain FC18, were obtained from Ilumina MiSeq paired-end. After, the raw sequences were assembled with SPAdes software, annotated in RAST and in Prokka pipeline.

Comprehensive analyses and differential genomic comparisons against *Rhodopirellula baltica* SH1^{T}, *Blastopirellula marina* DSM 3645 and *Planctomyces limnophilus* DSM 3776 were done after annotation. General features of the genomes were analysed, showing to be in the average of the other genomes from *Planctomycetales*. Genome mining with antiSMASH showed some interesting gene candidates with putative antimicrobial activity molecules. Furthermore, in the genome comparisons performed, LF1, UC8 and FC18 showed to possess unique proteins that can be connected with the microenvironment they were isolated from, the complex macroalgae biofilm.

Keywords: Planctomycetes, *Rubripirellula obstinate*, LF1, *Roseimaritima ulvae*, UC8, strain FC18, *Rhodopirellula baltica, Blastopirellula marina, Planctomyces limnophilus,* Bioinformatics, Sequencing, Assembly, Annotation, Genome, Antimicrobial activity, Biofilms.

Table of Contents

List of Abbreviations

®, ™	registered trademark
#	number
µL	microliter
µM	micromolar
µm	micrometer
16S rRNA	16S ribosomal ribonucleic Acid
ºC	Celsius
aa	aminoacids
ASW	artificial seawater
bp	base pairs
BLAST	Basic Local Alignment and Search Tool
CDS	coding DNA sequence
CECT	Spanish type culture collection
COG	Clusters of Orthologous Groups
ddH_2O	distilled water
DNA	deoxyribonucleic acid
dNTP	deoxynucleoside triphosphate
DSM	Deutsche Sammlung von Mikroorganismen (German Collection of Microorganisms)
e.g.	*exempli gratia* (for example)
et al.	*et alii* (and others)
gDNA	genomic DNA
G+C	guanine + cytosine
GO	gene ontology
HEPPSO	N-(2-Hydroxyethyl)piperazine-N′-(2-hydroxypropanesulfonic acid)
HGT	horizontal gene transfer
i.e.	*id est* (that is)
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LMG	bacterial culture collection from the department of Biochemistry and Microbiology, Faculty of Sciences of Ghent University
LGT	lateral gene transfer
LPS	leaf of lipopolysaccharide
LTS	long term support

LUCA	Last Universal Common Ancestor
Mbp	mega base pairs
M13	M13/607 medium
min	minute(s)
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NRPS	nonribosomal peptide synthases
OM	outer membrane
ORFs	open reading frames
ORI	origin of replication
PCR	polymerase chain reaction
PEG(s)	protein encoding gene(s) = CDS(s)
PG	peptidoglycan
PVC	*Planctomycetes* – *Verrucomicrobia* –*Chlamydia* superphylum
PKS	polyketide synthases
PRB	*Pirellula* – *Rhodopirellula* – *Blastopirellula* group
RAST	Rapid Annotation using Subsystem Technology
rDNA	ribosomal DNA
RNA	ribonucleic acid
rRNA	ribosomal RNA
s	second(s)
spp.	species
TBE	Tris-Borate-EDTA
tRNA	transfer RNA

Table Index

Figure Index

.

Part of this work was presented as an oral presentation in an international congress dedicated to the PVC superphylum bacteria "*Planctomycetes-Verrucomicrobia-Chlamydiae* Superphylum: New model organisms", 2-4 June 2015, Seville.

A manuscript of this work is under preparation to be submitted to Frontiers in Microbiology, in a special volume dedicated to the PVC conference.

1. Introduction

1.1 The phylum Planctomycetes

The Planctomycetes are a group of the domain *Bacteria* composed by members with peculiar and unique morphological, genetic, metabolic and physiological identity. They belong to the monophyletic *Planctomycetes*, *Verrucomicrobia* and *Chlamydiae* (PVC) superphylum (Wagner and Horn, 2006) and are a ubiquitous group of bacteria that are globally present in a myriad of ecosystems including aquatic and terrestrial habitats (Andrew et al., 2012; Winkelmann et al., 2010). Cultivation-independent techniques revealed their presence in extreme environments like hot springs (Tekere et al., 2013), glacial waters (Liu et al., 2006), acidophilic habitats (Lucheta et al., 2013), hydrocarbon polluted environments (Abed et al., 2011). Planctomycetes can also live in association with other organisms such as algae (Lage and Bondoso, 2011), sponges (Pimentel-Elardo et al., 2003), corals (Webster and Bourne, 2007), macrophytes (Hempel et al., 2008) and prawns (Fuerst et al., 1997). Besides, they play important roles in the biogeochemical cycles of the carbon (McCarren and DeLong, 2007), nitrogen (Kalyuzhnyi et al., 2010) and sulphur (Glöckner et al., 2003; Wegner et al., 2013).

Their unusual prokaryotic cell plan, peptidoglycanless proteinaceous cell wall, common budding reproduction as well as other previously mentioned characteristics, led to an increasing interest on these organisms, over the last decade. Despite the widely accepted cell plan theory for the cellular planctomycetes envelope proposed by John Fuerst (Fuerst, 2005) structural and genetic evidences gave support to a different concept in the last years (Santarella-Mellwig et al., 2013; Speth et al., 2012). Recently, Jeske et al., 2015 and van Teeseling et al., 2015 observed that *Planctomyces limnophilus* and *Kuenenia stuttagertensis,* respectively, possess peptidoglycan (PG) in their cell wall in a somehow Gram-negative cell wall structure. These observations gave incentive to the “planctomycetology” world to get new and more accurate information about this diverse and singular phylum.

The first planctomycete having its genome completely annotated was *Rhodopirellula baltica* SH1^T (Glöckner et al., 2003). By that time, it was the largest prokaryotic genome ever annotated. This annotation was an added value to get more information about this group (Teeling et al., 2004), yet very much unknown. Nowadays, despite

the wider knowledge already acquired due to the many studies done both *in silico* and *in situ*, many questions on the biology of *Planctomycetes* are still left to answer and to uncover.

1.2 Taxonomic and phylogenetic relevance

First observed by the Hungarian biologist Nandór Gimesi in 1924 (Gimesi, 1924; Langó, 2005), planctomycetes were initially thought to be a plancktonic fungus due to the morphologic similarities to the fungi. *Planctomyces bekefii* (Fig.1.1), possesses a unique phenotype, is widespread in many aquatic habitats but not yet cultivated in pure culture (Ward, 2010). The first isolation in pure culture of a planctomycete was only possible in 1972. This isolate was mistakenly identified as *Pasteuria ramosa* (Staley, 1973), a *Daphnia* parasite. Only in 1983, Starr *et al.* linked this species with the genus *Planctomyces*. Then *P. ramosa* was re-assigned as *Pirella staleyi* due to the differences of this strain (ATCC 27377^T) with species such as, *Planctomyces bekefii* Gimesi 1924 and *Planctomyces maris* strain ATCC 29201^T (Schlesner and Hirsch, 1984). Another re-assignment had to be done as the genus *Pirella* was already attributed to a fungal genus, leading into misunderstandings. Hence, from then on, strain ATCC 27377^T (Schlesner and Hirsch, 1987) became *Pirellula staleyi* (Schlesner and Hirsch, 1987) that together with other species like *Pirellula marina* are currently one of the existent taxonomic genera in Planctomycetes.

In 2001 the phylum *Planctomycetes* was recognised as a separate phylum (Garrity and Holt, 2001). Nowadays, according to the second edition of Bergley's Manual of Bacteriology (Garrity et al., 2005), this phylum comprises two classes – *Planctomycetia* (Ward, 2010) and *Phycisphaerae* (Fukunaga et al., 2009). The *Candidatus* "Brocadiales", a deep-branching planctomycete order capable of performing anammox oxidation (Jetten et al., 2010; Jetten, 1998), is grouped in the *Planctomycetia* class; nevertheless its affiliation is yet to be confirmed. Some authors disagree with this classification that has into consideration the molecular taxonomy and defend that the difference between these groups is phenotypically significant to consider the *Candidatus* "Brocadiales" a new class, instead of an order (Fuchsman et al., 2012; Fuerst, 2013). Based on the 16S rRNA gene, this phylum belongs to the PVC super phylum. The different groups of the PVC super phylum are also believed to share similar cell structure and plan. Despite some previous studies (Jenkins and Fuerst, 2001; Ward et al., 2000) in the beginning of this century that did not support

the PVC relationship, later it was indeed proved with molecular observations such as the analysis of the 16S rRNA gene, ribosomal proteins (Hou et al., 2008) and ribosomal DNA, rDNA (Wagner and Horn, 2006). Furthermore, the common phenotypic attributes, such as the supposedly lack of PG in the cell wall shared by both planctomycetes and chlamydiae, the compartmentalised cell plan and the presence of unusual coat proteins observed in planctomycetes and verrucomicrobia (Santarella-Mellwig et al., 2010) are other indicators supporting a close phylogenetic relationship. Besides *Planctomycetes*, *Verrucomicrobia, Chlamydiae* and *Lentisphaerae,* the PVC super phylum also comprehends the *Candidatus "*Poribacteria", OP3 and the candidate division WWE (Wagner and Horn, 2006). Planctomycetes are a heterogeneous group and have been considered either a deep branch (Schmid et al., 2003) within the domain *Bacteria* or a rapidly evolving group (Woese, 1987). However, its position within *Bacteria* is still under debate. Possessors of unique features, they have been of great help to scientists committed to better understand the evolution of cellular organization and its complexity. The close resemblance of many of the PVC features with the ones of Eukaryotes and Archaea include (1) the biosynthesis of sterols in *Gemmata obscuriglobus* (Pearson et al., 2003) important for the membrane fluidity and permeability, very typical in eukaryotic organisms; (2) presence of lipids found in eukaryotes like palmitic, oleic and palmitoleic and in archaeal organisms like the ether-linked lipids (Strous et al., 2002); (3) budding division, commonly typical of eukaryotes; (4) complex internal membranes and presence of membrane coats (Pilhofer et al., 2007); (5) fstZ and tubulin homologues, (Pilhofer et al., 2007); (6) existence of endocytosis-like process for protein uptake, well observed in *G. obscuriglobus* (Lonhienne et al., 2010). All these features, among others, seem to indicate that the PVC super phylum might be

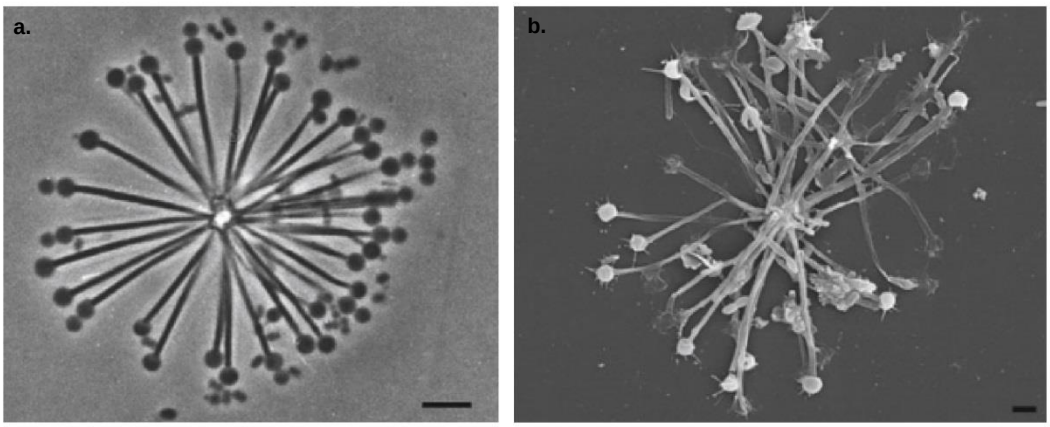


Fig. 1.1 - *Planctomyces bekefii* rosette a. Phase contrast micrograph of *P. bekefii* (Bar =5 µm) **b.** Scanning electron micrograph of *P. bekefii* (Bar = 1 µm). Adapted from Fuerst, 2013.

connected by a common ancestor. For example, Devos and Raymond data claimed that Planctomycetes should be considered candidates in the transition between the prokaryotes and eukaryotes, taking in consideration an existence of homology and evolutionary relationships between the PVC organisms (Devos and Reynaud, 2010; Reynaud and Devos, 2011). Some other scientists discard this hypothesis and believe that there might have been either horizontal gene transfer (HGT) or a convergent evolution to eukaryotic organisms (McInerney et al., 2011)

Other hypotheses and observations have been proposed looking at the conserved

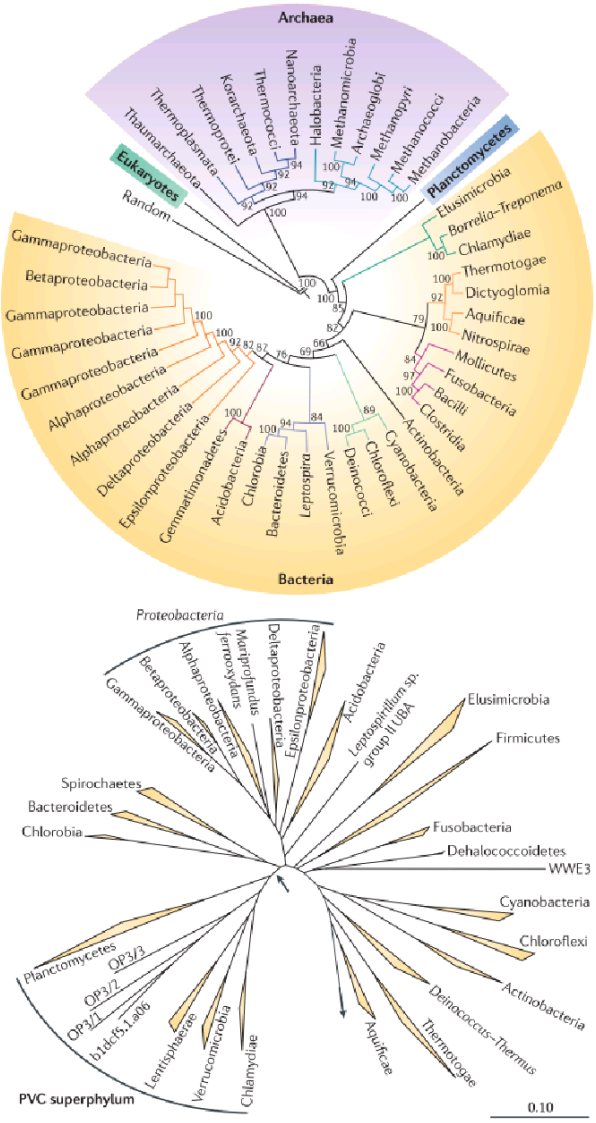


Fig. 1.2 - Phylogenetic relationship between planctomycetes and other organisms. a. Domains Bacteria, Archaea and Eukarya compared by the feature frequency profiles of whole proteomes, showing a deep-branching position for planctomycetes relative to other bacterial phyla. b. Phylogenetic tree based on the 23S rRNA gene representing the planctomycetes phylum and their relationship to other bacterial related phyla forming the PVC super phylum. Adapted from Fuerst and Sagulenko, 2011.

positions in the ribosomal rRNA, placing *Planctomycetes* as the first emerging bacterial group (Brochier and Philippe, 2002). Afterwards, they were considered not to be in the first line of divergence by Di Giulio (Di Giulio, 2003) who looked at different conserved positions at the ribosomal rRNA sequences defending that the ancestor of the domain Bacteria was most probably a hyperthermophile, instead of a mesophilic organism belonging to the Planctomycetales order. This observation was then supported by Barion et al. in 2007 analysing the phylogeny of twenty different concatenated proteins. After sequencing and comparing the first genomes from members belonging to the Planctomycetes, no great amount of genes shared with Archaea and Eukaryotes was observed, as it was firstly thought (Fuchsman and Rocap, 2006). Another study reached dissimilar observations, it defended that *Planctomycetes* are not that different from other bacteria having conserved functional roles and protein domains (Nasir et al., 2011). On the other hand, once again, the Planctomycetes were considered at the basal position of *Bacteria* when looking at the whole-proteome phylogeny of prokaryotes (Jun et al., 2010) (Fig. 1.2). This last hypothesis is currently, the most commonly agreed by the scientific community. Nonetheless, whether *Planctomycetes* have retention of a proto-eukaryotic LUCA or simply show a convergent evolution of eukaryotic-like features, the position of this phylum in the Tree of Life is still a controversial topic. This fact reinforces the importance of a more detailed study of the *Planctomycetes* and PVC organisms in order to have a clearer and more precise perspective in the real position they have.

1.3 Cell morphology and structure

The striking phenotypical traits and cell biology from many members of Planctomycetes are remarkable since they divide in general through an asymmetric buddy, without the presence of FtsZ, have a complex life cycle and comprise a complex cell plan uncommon in bacteria.

The most outstanding trait in a planctomycete is its internal organisation. They have been defined as possessors of a distinctive cell plan, presenting internal compartmentalization which divides the cytoplasm in two parts – the pirellulosome (containing the ribosomes) and the paryphoplasm (Lindsay et al., 2001). This compartmentalized cell plan is also present in the *Verrucomicobria* members (Lee et al., 2009). The anammox bacteria have an anammoxosome, where the anaerobic ammonium oxidation happens, and the genus *Gemmata has* a nucleoid surrounded by

two membranes, resembling a eukaryotic nucleus. However bacteria are non-compartmentalized organisms, lacking organization in organelles and with a dispersed DNA all over the cytoplasm. Therefore, the observation of the cell compartmentalization corroborates the hypothesis of the existence of a homologous relationship of Planctomycetes and eukaryotes aforementioned (Forterre and Gribaldo, 2010; Fuerst and Sagulenko, 2012).

Over the past years the complex internal membrane system has been studied and a new concept challenged the canonical idea of the cell structure of the Planctomycetes. Using *G. obscuriblogus* as reference organism and microscopic methods (electron tomography) to analyse the images of the internal membrane system complexity, it was observed that the compartmentalization so long widely accepted might be misleading. In fact, Santarella-Mellwig et al. (2010) observed a complex endomembrane system of cell membrane invaginations putting aside the idea of compartmentalization. In a study with different species, Lage et al. (2013) reported a similar observation. If these observations are indeed correct, it means that there is no paryphoplasm but a common bacterial periplasm with a flexible cytoplasmic membrane prone to invagination. In order to detect if these observations were correct, the presence of Gram-negative outer membrane (OM) biomarkers present in the genome of *Planctomycetes* and *Verrucomcrobia* available in Genbank database were assessed (Speth et al., 2012). The results obtained supported the last observations.

The cell division ring, normally located in the cytoplasm and also typical of the Gram-negative bacteria was also previously observed in the paryphoplasm of "*Candidatus Kuenenia stuttgartiensis"* (van Niftrik et al., 2010). Furthermore, traits indicative of the presence of a Gram-negative cell wall structure in planctomycetes have been observed. The existence of an outer-membrane like structure was evidenced by the presence of unusual glycolipids that are normally part of the lipopolysaccharide (LPS) a structure present in the outside leaf of the outer membrane (OM) (Lugtenberg and Van Alphen, 1983). Due to the clear characteristics of the OM, more specifically, some genes related with the biosynthesis of lipid-A (an important LPS components) and 2-keto 3-deoxy-D-manno-octulosonate (KDO) were also found in the genomes of *Planctomycetes* and *Verrucomicrobia* (Glöckner et al., 2003; Sutcliffe, 2010). These observations are the key to approach Planctomycetes towards a more Gram-negative-like cell plan.

Since the 80's due to the great resistance of Planctomycetes to beta-lactam antibiotics like ampicilin, that target peptidoglycan synthesis, scientists verily believed that

Planctomycetes lacked PG (Liesack et al., 1986) having a proteinaceous cell wall instead (Fuerst and Sagulenko, 2011; Giovannoni et al., 1987). Despite these widely acknowledged traits of Planctomycetes and of many organisms in the PVC group, recent reports have shown contradictory observations. The anammox planctomycetes showed sensitivity towards ampicillin and lysozyme suggesting a presence of peptidoglycan-like components in their cell wall (Hu et al., 2013). Chlamydiae was also found to produce PG, a component they were equally thought to lack (Liechti et al., 2014). Glöckner et al. (2003) analysed the complete genome sequence of *R. baltica* where it was noticed the presence of some genes involved in the formation of N-acetyl-D-glucosamine, a monomer of peptidoglycan. However, as the key enzymes were lacking, little or no importance was given to this observation and it was proposed a possible loss of these genes, followed by the development of a proteinaceous cell envelope over time (Glöckner et al., 2003). Two very recent studies revealed the presence of PG in *P. limnophilus* (Jeske et al., 2015) and in the anammox bacterium *Kuenenia stuttgatiensis* (van Teeseling et al., 2015). Therefore, Planctomycetes are not an exception to the presence of PG among bacteria. This presence was not only observed *in silico*, doing bioinformatics analysis, but also *in situ*, by the biochemical and microscopic analyses of the planctomycetal cell wall (Jeske et al., 2015; van Teeseling et al., 2015). These studies suggest and support the perspective that Planctomycetes might have a potentially ancient molecular mechanism of cell division (Leaver et al., 2009) not associated with the presence of FtsZ.

Several members of the planctomycetes divide by budding reproduction. Budding reproduction is rarely observed in bacteria which normally divide by binary fission with the help of a GTPase FtsZ protein. However, some planctomycetes divide by binary fission like the case of *Phycisphaera mikurensis* (Fukunaga et al., 2009) and the anammox bacteria (van Niftrik et al., 2009).

Phenotypically, planctomycetes cell shapes vary from pear shaped, spherical to ovoid and many times the cells have polarity: a vegetative pole with a flagellum/stalk (Fuerst, 1995) and a reproductive pole. The cell wall can harbour crateriform structures and fimbriae (Starr et al., 1983). One feature that is very known and correlated with planctomycetes, specially the *Pirellula – Rhodopirellula - Blastopirellula* (PRB) group, is the formation of rosettes (Fig. 1.3.). Some organisms are unicellular as well as filamentous (Ward et al., 2006). Planctomycetes may display motility in the first stage of their life cycle or display gliding motility as *Isosphaera pallida* (Giovannoni et al., 1987). Planctomycetes also easily attach to other cells or surfaces by the loss of their

flagellum and by the production of a holdfast substance, yet to be characterised (Lage and Bondoso, 2012).

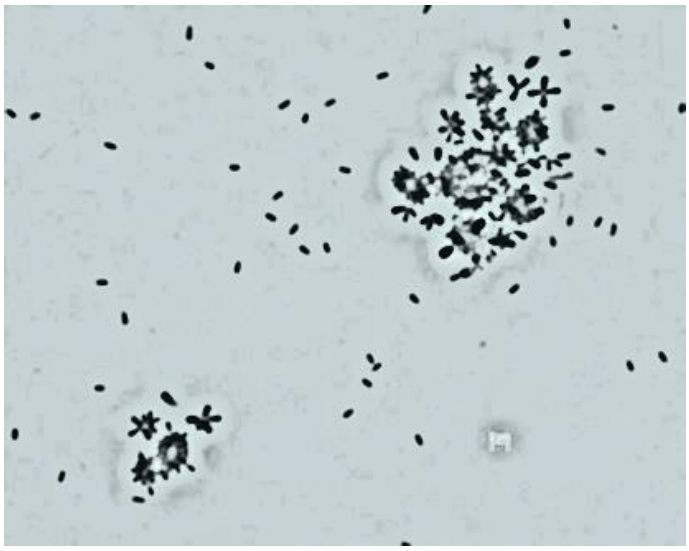


Fig. 1.3 – *Rhodopirellula rubra* strain LF2 observed in optical microscopy. Observation of the typical rosettes from the PRB group. Image gently provided by Olga Lage.

1.4 Physiology and environmental relevance

Due to a great metabolic diversity, Planctomycetes are considered to have a very important role in the global environmental cycles, contributing to the global carbon (Glöckner et al., 2003) nitrogen and sulphur cycles. For the nitrogen cycle, all the anammox of the candidate order Brocadiales are important as they perform the anaerobic ammonium oxidation, a unique pathway that converts ammonium to nitrogen in an oxygen independent way (Strous et al., 1999, 2002). These organisms have been used successfully in industry with a biotechnological application in the wastewater treatment plants (Kartal et al., 2010) and in the utilization of some enzymes for specific biotechnology processes (Sheldon, 2011). Apart from these chemolitothropic anammox planctomycetes, all the others are chemoheterotrophs with carbohydrates as primary source of carbon. A great majority of them are mesophilic but they also can be thermomophiles such as *I. pallida* isolated from the North American hot springs (Giovannoni et al., 1987). Other extreme locations where planctomycetes have been isolated are the acidic wetlands (Kulichevskaya et al., 2007, 2008, 2012), hypersaline water (Schlesner, 1989), hydrocarbon polluted environments (Abed et al., 2011) and

other polluted habitats (Chouari et al., 2003). With these observations it can be suggested that some Planctomycetes may have a role in the degradation of hydrocarbons and other pollutants. Some planctomycetes were also reported to resist to high concentrations of inorganic nitrogen compounds and to sodium azide even though it affected the cell respiration till a certain extent (Flores et al., 2014). Hence, it was shown that the planctomycetes are possibly good candidates for assessing water quality (Flores et al., 2014). Planctomycetes are ubiquitously spread in the environment, including in association with eukaryotic hosts like invertebrates (Fuerst et al., 1997), corals (Webster and Bourne, 2007), sponges (Pimentel-Elardo et al., 2003; Webster et al., 2001), prawns (Fuerst et al., 1997), macrophytes (Hempel et al., 2008), living the rizosphere of plants (Zhang et al., 2013, 2010) as well as associated living in the gut microbiome of humans (Cayrou et al., 2013). Furthermore, some studies reveal planctomycetes as being dominantly present in macoalgae and biofilms (Bondoso et al., 2011, 2014a; Lage and Bondoso, 2014). The kelp *Laminaria hyperborea* has an abundance of planctomycetes up to 51% in the bacterial community of their biofilm (Bengtsson and Øvreås, 2010). Many other species of macroalgae have planctomycetes living in association with them (Lage and Bondoso, 2014). In fact, algae harbour a high diversity of microbial communities (as well as other organisms) that benefit from a wide variety of organic carbon sources produced by them. Microbial communities are as well believed to have an important role in the host's metabolism, development and defence (Armstrong et al., 2001). Green algae like *Ulva australis* (Longford et al., 2007; Tujula et al., 2010), *Ulva prolifera* (Liu et al., 2010), *Ulva compressa* (Hengst et al., 2010), *Ulva* sp. and *Ulva intestinalis* (Lage and Bondoso, 2011) have planctomycetes frequently colonizing their surfaces. Isolates have also been obtained, among others, from red algae *Phorphyra dioca, Chondrus crispus, Gracilaria turuturu* (Lage and Bondoso, 2011), *Laurencia dendroidea* (de Oliveira et al., 2012), *Delisea pulchra* (Longford et al., 2007). Planctomycetes isolation has also been obtained from the brown macroalgae *Fucus spiralis*, *Laminaria sp*., *Sargassum muticum* (Lage and Bondoso, 2011) and *Laminaria hyperborean* (Bengtsson and Øvreås, 2010). Lachnit et al., 2011 observed that the planctomycetes associated with red and brown macroalgae showed higher diversity than the ones colonizing green algae. Despite being very abundant, Planctomycetes are not generally the major group found in algae. The two major are *Bacteroidetes* and *Proteobacteria* followed by *Firmicutes, Actinobacteria, Verrucomicrobia*, and, only then *Planctomycet*es (Goecke et al., 2013) as shown in Fig. 1.4. Currently, more than 60 planctomycetes' different

Operational Taxonomic Units (OTUs) were reported to be associated with macro algae being only around 10 species isolated in pure culture (Bengtsson and Øvreås, 2010; Bondoso et al., 2011, 2015; Winkelmann and Harder, 2009). One very interesting observation in the planctomycetes associated with macroalgae is that they seem to have the capacity to adapt to microenvironments created by molecules released by macroalgae. The macromolecules include sulfated polysaccharides such as carrageenan, agar, alginate, fucan, laminarina, cellulose and ulvan (Lage and Bondoso, 2014). Glöckner et al. (2003) analysing *R. baltica* SH1^{T} complete genome, reported the presence of 110 different sulfatase genes. These can be related with energy and carbo requirements from sulphated compounds.

Different *Rhodopirellula* strains have shown sulphatases genes and sulphatase expression profiles in bacteria cultured in distinct sulphated polysaccharides (Wegner et al., 2013). Polysaccharide uptake was also observed by the utilization of some

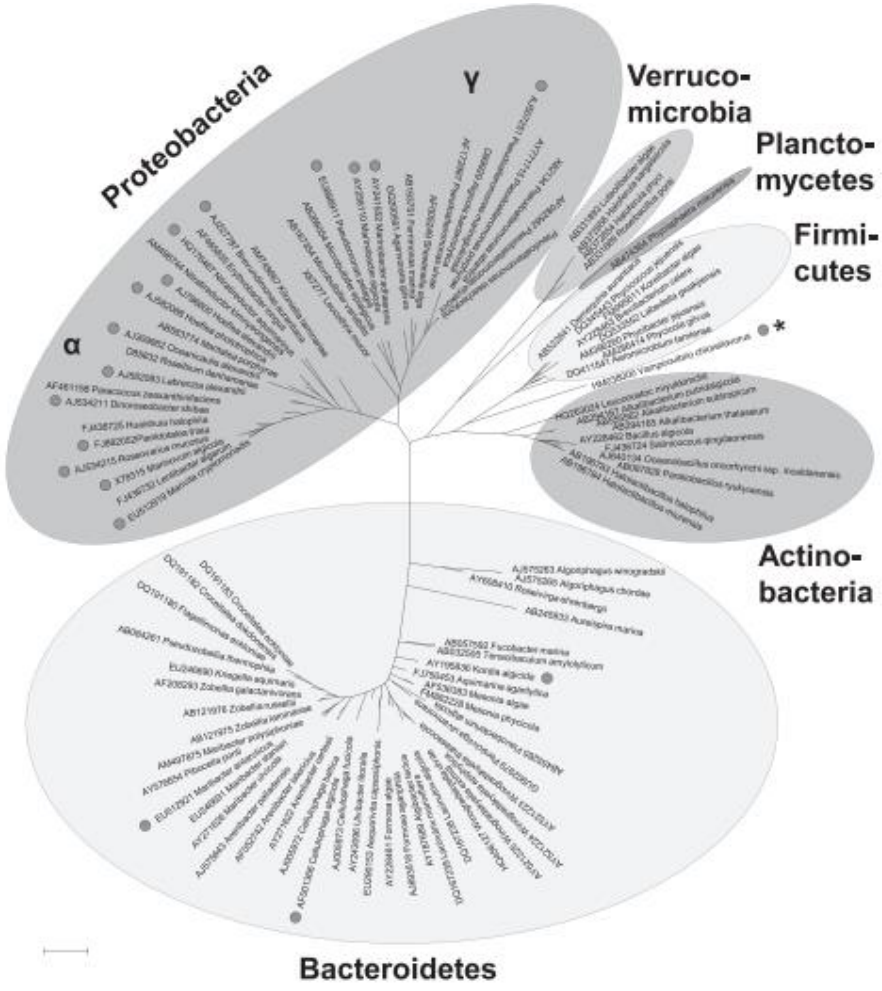


Fig. 1.4 - Maximum Likelihood phylogenetic tree of the 16S rRNA gene sequences from bacterial species isolated from algae. This figure shows the most dominant phyla found in macroalagae. The scale bar indicates 0.1 change per nucleotide. Adapted from Goecke et al. 2013 (Goecke et al., 2013).

polymers such as N-acetylgalactosamine, mannitol, D-glucoronic acid, pectin and laminarin (Jeske et al., 2013). *R. rubra* and *R. lusitana* isolated from macroalgae showed to uptake great part of the monomers constituting the main polysaccharides released by marcoalagae (Bondoso et al., 2015).

1.4 Secondary metabolites

Planctomycetes are possessors of large genomes, average 6.9 Mbp (Jeske et al., 2013) and have intricate life cycles. These traits are normally typical of bacteria that are known to be potential producers of bioactive compounds like *Actinobacteria* and *Myxobacteria*, suggesting that the planctomycetes may have the capacity to produce these compounds.

Jeske et al. (2013) performed genome mining with 13 planctomycetes genomes and found a high number of clusters and genes for the production of secondary metabolites like bacteriocin and putative lantibiotic enconding genes among others. Nevertheless, specific environmental conditions are needed for many of the genes to be expressed (Jeske et al., 2013). A similar study performed in *R. baltica* showed the presence nonribosomal peptide synthethases and polyketide synthases genes that synthesize enzymes involved in the synthesis of five different, unknown bioactive products (Donadio et al., 2007). Planctomycetes associated with algae may secrete secondary metabolites such as growth factors or antimicrobial compounds that may benefit algae. These should also be important for the planctomycetes during algae colonization and in their defence against competitors. Other observation, supporting the secondary metabolites' production by *Planctomycetes* is the relationship between planctomycetes and algae which can make them to use algae's excreted compounds to trigger the production of those compounds (Jeske et al., 2013). These recent observations may lead to the discovery of drug production with antimicrobial activity by *Planctomycetes*.

1.6 Genomics and Bioinformatics of prokaryotes

The first bacterial genome sequences ever published happened 20 years ago (Fleischmann et al., 1995; Land et al., 2015). Sequencing techniques have been developing at a rapid pace and it is every time easier and more affordable to sequence genomic DNA (gDNA). The study of the diversity of microorganisms present in the environment was for many years mainly done with culture dependent methods.

However, as evidenced by the “Great Plate Count Anomaly” (Staley and Konopka, 1985), in marine environment, not even 1% of the organisms are able to be cultured and accessible. This fact is a reminder of how little it was, and, still is known about the microbial diversity yet to be discovered. With the elucidation of the structure of the DNA and its identification as the structure that harbours the genetic information of organisms from all domains of life in 1953 (Watson and Crick, 1953) an important step towards the understanding of life in general was created. The study and identification of the small subunit of the ribosomal RNA (16S rRNA) as a tool to identify microbial diversity (Woese, 1987), allowed taxonomic assignment and phylogenetic trees constructions. Furthermore, the advances in DNA sequencing technologies like 454 pyrosequencing techniques and others known as “culture independent methods” made it possible for researchers to analyse with more detail the general diversity, the genetic potential and determine abundant members of whole bacterial communities (Giovannoni et al., 1990). The still widely used approaches based on 16S rRNA gene allow reliable information on bacterial family and genus, but reveals poor resolution at species level (Case et al., 2007). With the development of the genomic era this single gene comparison approach is being replaced by more cyclopaedic approaches, able to thoroughly analyse, compare and classify a myriad of genomes at a time. The metagenomic data, covering all DNA present in a sample, allowed to have detailed overviews on numerous environmental, human and animal microbiomes (Land et al., 2015).

There has been an exponential number of sequenced genomes growing over the last two decades in the databases (Fig. 1.5) and the rapid increase of information has led to an overflow of data. Hence, lots of tools are constantly appearing making genome sequencing cheaper and abundant over time. Bioinformatics, nowadays, plays an important role in decoding prokaryotic lifestyles, relationships and contributing with novel insights towards diversity.

There are three generations of sequencing so far. Most genomes were sequenced by the Sanger method which made draft genomes harder to be completed and expensive. The second generation, the “Next-generation sequencing” (NGS) produces shorter reads, providing means for rapid and high throughput sequencing, data generation at low cost increasing the coverage needed making thus easier for the genome to be closed. However, the *de novo* assembly i.e. assembly done for the very first time with no prior knowledge of the genome, is very hard to obtain without a proper scaffold or paired-end reads. Sequencing based on Illumina technology was reported as cost-

effective in order to generate draft genomes from microbial organisms without a significant loss of information (Mavromatis et al., 2012).

The third generation sequencing is the single-molecule sequencing and it is able to produce several thousand base paired reads (Land et al., 2015).

Looking at the genomes it was observed that there are a lot of redundancies in gene replications and that even within species the degree of genetic variation can sometimes be large (Binnewies et al., 2006). Genome comparisons, the study of groups of conserved proteins and proteomes enable an analysis with a bigger scale being able to infer a more accurate phylogenetic profiling, related functional pathways and resolve taxonomic enigmas. The more genes considered the better the taxonomic resolution and the lesser the sensitivity to horizontal gene transfer (Oren and Papke, 2010)

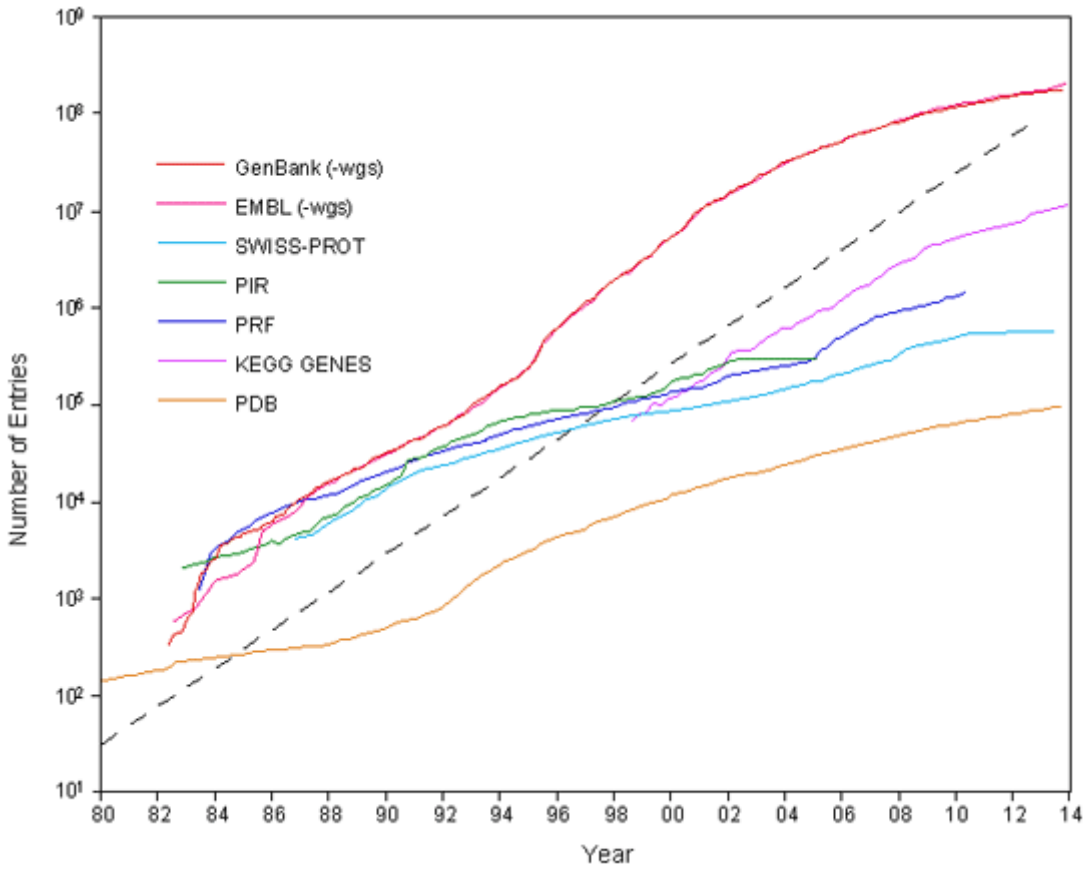


Fig. 1.5 - DNA sequencing data growth over the past 30 years. Representation of the growth of sequence and 3D structure databases. (http://www.kanehisa.jp/en/db_growth.html)

1.6.1 Genome assembly and genomes belonging to Planctomycetes

Rhodopirellula baltica SH1^{T} was the very first genome belonging to the *Planctomycetes* being sequenced in 2003 (Glöckner et al., 2003) . From that moment on, some planctomycetes genomes have been sequenced completely or partially, resulting in a closed or draft genome respectively. Up to date there 17 sequenced genomes, either draft or complete, of plantomycetes strains (Table 1.1).

Table 1.1 - Planctomycetes strains with their genome sequenced.

Sequenced Planctomycetes
Blastopirellula marina DSM 3645
Gemmata obscuriglobus UQM 2246
Isosphaera pallida ATCC 43644
Phycisphaera mikurensis NBRC 102666
Pirellula staleyi DSM 6068
Planctomyces brasiliensis ATCC 49424
Planctomyces limnophilus DSM 3776
Planctomyces maris DSM 8797
Planctomycete KSU-1
Rhodopirellula baltica SH 1
Rhodopirellula europaea 6C
Rhodopirellula maiorica SM 1
Rhodopirellula sallentina SM 41
Rhodopirellula sp. SWK7
Schlesneria paludicola DSM 18645
Singulisphaera acidiphila DSM 18658
Zavarzinella formosa DSM 19928

1.7 Aim of the dissertation

The aim of this dissertation is focused on the sequencing and subsequent analysis of three genomes of strains belonging to *Planctomycetaceae* family – *Rubripirellula obstinata* strain LF1, *Roseimaritima ulvae*, strain UC8, and a yet uncharacterised strain FC18. Furthermore, the present project intends to provide novel insights into the characterisation, metabolism and lifestyle diversity of these strains, proposing findings related to their environmental conditions and also, enlarge the knowledge in the *Planctomycetes* phylum.

2. Materials and Methods

2.1 Biological material

The plancomycetes strains used in the current study were stains isolated from macroalgae biofilm (1) *Ulva sp*. sampled in Carreço (41º44'N, 8º52'W) (2) *Laminaria sp*. in Porto (41º19'N, 8º40'W) and (3) *Fucus spiralis* from Carreço (41º44'N, 8º52'W), respectively *Roseimaritima ulvae* strain UC8 (Fig. 2.1 b) (Bondoso et al., 2015; Lage and Bondoso, 2011, GenBank: HQ845508.1), *Rubripirellula obstinata* strain LF1 (Bondoso et al., 2015; Lage and Bondoso, 2011, GenBank: DQ986201.2) (Fig. 2.1 c) and an uncharacterised strain FC18 (Lage and Bondoso, 2011, GenBank: HQ845450.1) (Fig. 2.1 a). An example of the growth of strains on top of macroalgae portions during isolation is shown in Figure 2.1 d. Fig. 2.2 shows the phylogenetic relationship of these strains with other related planctomycetes.

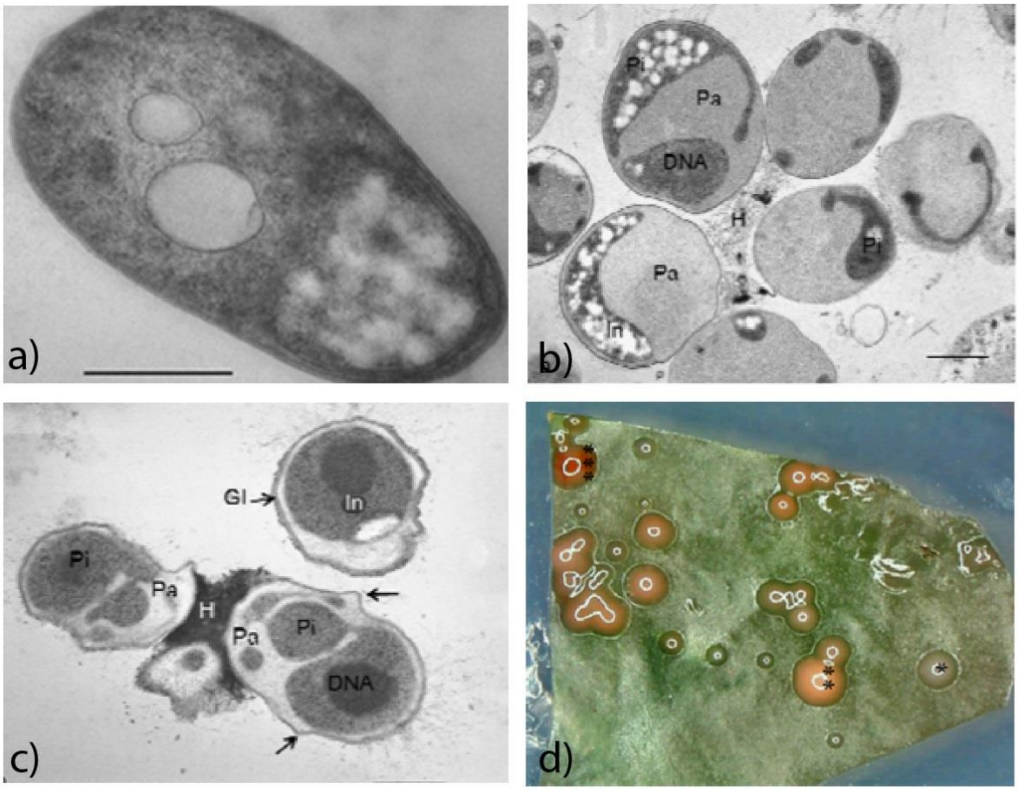


Fig. 2.1 - Transmission electron microscopy (TEM) of the strains under study and planctomycete colonies isolated from an algal surface. a) TEM of strain FC18 b) TEM of strain UC8 c) TEM of strain LF1 d) Colonies of Planctomycetes growing on the surface of a portion of *Ulva sp.* (adapted from Lage and Bondoso, 2011).

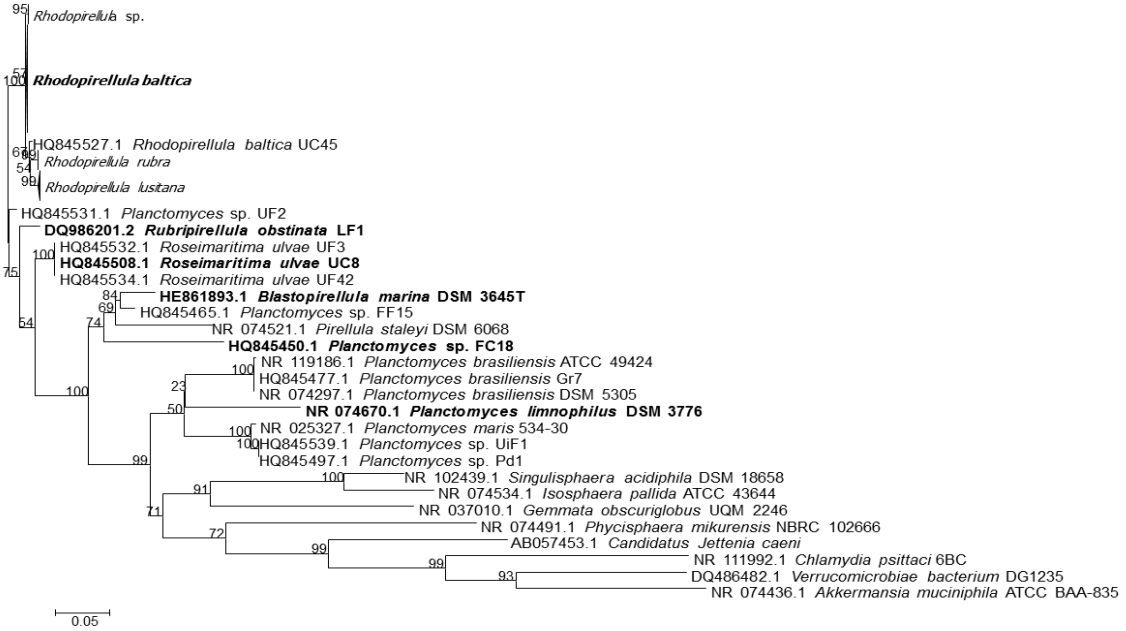


Fig. 2.2 - Phylogenetic 16S rRNA gene tree. It was generated by maximum-likelihood analysis and based on the General Time reversible model indicating the relationship of the strains under study to members of the Planctomycetes. The *Verrucomicrobia* bacteria were used as an outgroup. The numbers beside nodes are the percentages for bootstrap analyses. Scale bar = 0.05 substitutions per 100 nucleotides.

2.2 Genomic DNA extraction and 16S rRNA gene amplification and analysis

Genomic DNA from the three different batch cultures cultured in modified solid M13 (Lage and Bondoso, 2011) at 24 ºC was extracted in duplicate using the E.Z.N.A. ® Genomic DNA Isolation Kit, Omega Bio-Tek, VWR. In order to confirm the identity of the species under study the 16S rRNA gene was amplified: 1µL of the extracted gDNA, cooled on ice with 2 µM of the universal primers 27F and 1492r (Lane, 1991) in 25 µl of a PCR mixture (1x PCR buffer; 1.5 mM $MgCl_2$; 1 unit of GoTaq Flexi DNA Polymerase; 200 µM of each deoxynucleoside triphosphate (dNTPs)). The PCR program was performed in a MyCyclerTM Thermo Cycler (Bio-Rad) and consisted in an initial denaturing step of 5 min at 95 ºC; 30 cycles of 1 min at 94 ºC; 1 min at 52 ºC; and 90 s at 72 ºC; and a final extension of 5 min at 72 ºC. PCR products (5 µL) were visualized after electrophoresis in a 1.2 % agarose gel in 1X TBE buffer. The PCR products were sent to Macrogen to be purified and sequenced

After being sequenced, the 16S rRNA gene sequences received from Macrogen were cleaned in Chromas 2.12 software (http://technelysium.com.au/). ProSeq v 2.91 (Filatov D. A., 2009) was used to construct the consensus sequences of each strain. Consensus sequences were then blasted in GenBank, to confirm their identity.

2.2.1 gDNA quantification

To be able to know if the extracted gDNA had 2 microgram of high molecular weight DNA (requested for the Illumina sequencing) a gel electrophoresis was performed with the same aforementioned characteristics, mixtured with 2 µL of loading dye. Five microliters of GeneRuler DNA ladder Mix #SMO331/2/3 were used in the gel. Afterwards, the weight quantification was performed in a GenoPlex chamber (VWR) with the assistance of GenoSoft software (VWR). After the confirmation of the molecular weight, 95 µL of gDNA solution from the three bacterial strains were sent to the sequencing center in Cologne.

2.3 Genomic DNA sequencing

Genomic DNA sequencing was performed at the Genome Center of the Max Planck-Institute for Plant Breeding Research in Cologne, Germany. The Illumina MiSeq technology was the Next-Generation Sequencing technology platform used for sequencing. The genomic library preparation was performed with the NEB NextUltra™ DNA Library Prep Kit for Illumina, NEB. The Illumina method was performed in two (FC18 and LF1) or three (UC8) runs, generating 250 bp long paired-end reads downloaded from the local database of the Genome Centre in fastq format.

2.4 Genome assembly approaches

2.4.1 Raw data and assembly of "1st generation"

The 1st generation assemblies, *i.e.:* assemblies that are primarily done were manipulated in a UNIX system with Linux version 10.04.4 LTS 10, controlled with Bash language for programming in a command line console. The scripts and programs are provided in majority in python or perl language. All tools and software needed to use are presented in Table 2.1. In Table 2.2 are presented the three different approaches done with the assemblers, explained below.

6697558 paired-end reads from FC18, 6437529 paired-end reads from UC8 and 6856066 paired-end reads from LF1 of 250 bp length and of high quality were, separately, dynamically trimmed with SolexaQA v.2.2 (Cox et al., 2010). The 5' and 3' ends based on the quality values of the corresponding nucleotide positions were trimmed with DynamicTrim (trimming value of 10). Reads smaller than the mean length were discarded with LengthShort. This first step increased more confidence in the results of the assemblies.

Table 2.1 – Software and tools used for preparation, assembly and post-analyses of the sequence reads during assembly.

Software / Tool	Version	Objective	Source
DynamicTrim	SolexaQA v 2.2 package	Trim raw reads	Cox et al., 2010
LengthShort		Remove short reads	
Interleave		Concatenate paired-end read files	
Normalize-by-median	Khmer package v 1.0	Decrease redundancy	Brown et al., 2014
Filter-abund		Unsystematic coverage	
Extract-paired-reads		separation of paired and orphan files	
Bbmerge	v 4.0	Merge paired reads into single reads by overlap detection	http://bbmap.sourceforge.net/
Bbmap	v 32.x	Short-read aligner for DNA and RNA & mapping	
Bbtrim		Perform trimming and/or kmer-trimming on reads	
VelvetOptimiser	v 2.2.5	Genome assembly (1st generation)	Zerbino and Birney, 2008
IDBA-UD	v 1.1.1		Peng et al., 2012
SPAdes	v3.1.0 – Linux		Bankevich et al., 2012
Sequencher	v 4.6	Genome assembly (2nd generation)	
Geneious	R8	Organization and analysis of sequence data	Kearse et al., 2012

After the trimming step, the paired-end reads were normalized by the Khmer 1.0 (Brown et al., 2014), reducing redundancy. The normalized data was then ready to generate optimal assemblies with good contig length and N50 values, which are a qualitative parameter for evaluation of the assembly quality (Mäkinen et al., 2012). Split into paired-end and single-end reads they were then assembled with VelvetOptimiser v 2.2.5 (Zerbino and Birney, 2008) and SPAdes v 3.1.0 (Bankevich et al., 2012). For the

assembly using IDBA-UD v 1.1.0 (Peng et al., 2012) the steps were very similar to the ones mentioned above, with an initial trimming step. However, in this case it was not necessary to normalize the data, going straight to the assembly step. The output created from this first generation assemblies had a different number of units structured from overlapped region of the reads during assembly process, called contigs. These contigs represented a portion of the genomic sequence of the organism.

Table 2.2 – Assembly approaches used. This table shows the software necessary to obtain the assembly from each assembler. Each column represents one different approach and shows the intermediate steps.

Steps	**Tools**		
	Bbtools with bbmap_package	*Digital_normalization*	*Bbmap_idba*
Trimming	Bbduck.sh Adapter_trim Quality_trim	SolexaQA Dinamically Trim Filter by lenght	Bbmerge merging Bbtrim Adapter_trim
Normalization	Bbnorm.sh Normalization Error connection	khmer Interleave Add single reads Normalization	-
Assembler	SPAdes & Velvet		IDBA-UD

2.4.2 Assembly of "2nd generation"

The contigs obtained from the assemblies of 1st generation of each strain were put together and *de novo* assembled in Sequencher v 4.6 (Gene Codes Corporation, Ann Harbor, USA) in order to be possible to obtain fewer and longer contigs. Afterwards, to remove possible duplications in the contigs, reads were mapped onto contigs with GENEious R8 (Biomatters, Auckland, New Zealand) (Kearse et al., 2012) to identify possible contig elongations. The contigs used in the mapping were the two or three longest contigs obtained in Sequencher 4.6 output.

2.5 Quality check

To be aware of the quality of the assemblies, three different software types were used in distinct times. Both Metawatt v2.1 (Strous et al., 2012) and Quast v2.2 (Gurevich et al., 2013) were used after the assemblies. CheckM v0.9 (Parks et al., 2015), a newly launched software, was used. Quast analyses quantitatively the number of contigs,

N50, among other parameters. Metawatt and CheckM, alternatively, were able to detect contaminated sequences reducing thus, the number of contaminations.

2.6 Bioproject and Biosample submission

In order to inform the scientific community about this work, a submission in BioProject and BioSample projects belonging to the NCBI - National Centre for Biotechnology Information was done. BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium of coordinating organizations and collects data to provide users with an entry point into diverse data types. BioSample is a description of the biological source materials used in experimental assays, explaining many features of these materials.

Table 2.3 – Information required in the BioProject submission in NCBI.

Field name	FC18 (uncharacterised)	*Roseimaritima ulvae* UC8	*Rubripirellula obstinata* LF1
Organism domain	Bacterial	Bacterial	Bacterial
Phylogeny	Planctomycetes	Planctomycetes	Planctomycetes
Genus	-	*Roseimaritima*	*Rubripirellula*
Species	-	*Roseimaritima ulvae*	*Rubripirellula obstinata*
Strain	FC18	UC8	LF1
Geographic location	Carreço	Carreço	Porto
Latitude	41°44’ N	41°44’ N	41º09'N
Longitude	08°52’ W	08°52’ W	08º40'W
Depth	Zero	Zero	Zero
Altitude	Zero	Zero	Zero
Time of sample collection	2006	2006	2005
Ecosystem	Environmental	Environmental	Environmental
Habitat	Marine	Marine	Marine
Biotic relationship	Macro-algae association	Macro-algae association	Macro-algae association
Relationship to Oxigen	Aerobic	Aerobic	Aerobic
Isolation and growth conditions	25° C; M13 media or M600 isolated in HEPPSO buffered M13 in darkness	30° C; M13 with ASW 20-25% of salinity and 7.5 pH	25° C; M13 with minimum salinity of 50% and 7.5 pH
Temperature range	Mesophile	Mesophile	Mesophile

Field name	FC18 (uncharacterised)	*Roseimaritima ulvae* UC8	*Rubripirellula obstinata* LF1
Energy sources	Heterotroph	Heterotroph	Heterotroph
Source material identifiers		DSM 25454 = LMG 27778	LMG 27779 = CECT 8602
Sequencing method		Illumina MiSeq 2x250	
Assembly		SPAdes (IDBA-UD, Velvet)	

2.7 Automatic annotation

2.7.1 Rapid Automatic Annotation

The gemonic contigs belonging to the three strains were submitted in Rapid Annotation Subsystem Technology (RAST) v 2.0 (Aziz et al., 2008). This is an automated web service that allows a thorough analysis of the genome and identifies protein-encoding, rRNA and tRNA genes. It assigns functions to the genes with all the acquired information, it reconstructs metabolic pathways. Besides, it allows a comparative analysis with other annotated genomes in SEED-viewer. RAST, uses a "Highest Confidence First" assignment propagation strategy based on manually curated subsystems and subsystem-based protein families that automatically guarantees a high degree of assignment consistency (Aziz et al., 2008; Devoid et al., 2013) .These subsystems are groups of proteins related by function linked to a biological or structural process (Overbeek et al., 2005).

2.7.1.1 Comparison tools in SEED-viewer

In RAST the curated subsystems obtained are connected to a set of freely available group of protein families, known as FIGfams (Meyer et al., 2009), being, thus, the core component of RAST. As a compliment to RAST, SEED-viewer allows read-only access to the latest curated data sets, providing a numerous number of tools that allow comparisons between genomes, private or public.

2.7.1.1.1 Function based comparison tool

Function-based comparison tool allows comparing the genome under study with others in the database, associated with a complete subsystem. The table is sortable,

searchable by subsystem category or name, and downloadable. Annotations are assigned based on sequence similarity, while inclusion in subsystems is based on the annotation that matches a functional role of a subsystem. In this study, the genome annotations from UC8, LF1 and FC18 were compared between each other and also with *Rhodopirellula baltica* SH1^T and *Blastopirellula marina* DSM 3645 (for FC18 the latter).

2.7.1.1.2 Sequence based comparison tool

This SEED – viewer tool enables to select the genomes under study and assess the protein similarities. *R. baltica*, *B. marina*, LF1, UC8 and FC18 were used as reference genomes and compared with the three of LF1, UC8 and FC18.

2.7.2 Prokka

In order to be able to perform gene prediction and 1st annotation before the differential analyses, Prokka v1.11 software (Seemann, 2014) was also used. This second annotation is an on demand line software tool using UNIX system with CentOS Linux 6, controlled with Bash language for programming in a command line console. Prokka coordinates a suite of existing software tools such as BLAST+, HMMER, Aragorn, tbl2asn to achieve a rich and reliable annotation of genomic bacterial sequences. To detect the Open Reading Frames (ORFs) Prodigal software in the Prokka pipeline was used.

2.8 Homolog protein clusters – Differential analysis

With the objective of identifying protein clusters belonging to LF1, UC8 and FC18, OrthoMCL v 2.0 was the software chosen. OrthoMCL allows the analysis of the orthologs, paralogs and coortologs groups (Fischer et al., 2011). This analysis is based on protein sequences due to its higher sensitivity when comparing it to genomic sequences. It requires many steps and it was manipulated in a UNIX system with CentOS Linux 6, controlled with Bash language for programming in a command line console. In general, the steps pass through an all-vs-all BLAST, then an OrthoMCL Pairs program which makes the pairs or directory and at last, the MCL program that creates the clusters between the pairs. The threshold for a reciprocal best match was a BLAST result between two genomes with an expectation value E of less than 1e-5.

For the comparison, the genomes used were *R. baltica, B. marina*, FC18, UC8, LF1 and *Planctomyces limnophilus* DSM 3776. This latter strain was added to give a broader spectrum and more accurate data. After the results, it was possible to detect the genes in the clusters shared between each of the bacteria, using a specific script gently provided by Nicola Bordin. Also, it was possible to assess the shared and unique genes of LF1, FC18 and UC8. After the OrthoMCL output, three scripts were used: one for obtaining all the clusters in common in a 6 bacteria (LF1, UC8, FC18, *R. baltica*, *B.marina* and *P.limnophilus*) all-vs-all comparison, one for obtaining all the clusters in a 3vs3 (LF1, UC8, FC18) all-vs-all comparison and a last script that, provided with the results from the previous two scripts, extracted the FASTA sequences for each of the clustered proteins.

2.9.1 Protein identification

After assessing the common proteins between LF1, UC8 and FC18, the ones that resulted from the two reference genomes (*R. baltica and P. limnophilus*) were retrieved using UniProt database using the ID Retrieve/Mapping Tool (http://www.uniprot.org/uploadlists/) and one was retrived from GenBank (*B.marina*) (http://www.ncbi.nlm.nih.gov/genbank/). For the identification of the proteins shared between FC18, LF1 and UC8 (not belonging to any database) it was used InterProScan (Jones et al., 2014) and PSI-BLAST. The results were manually curated. In order to have more fruitful and accurate results a deeper analysis of the hypothetical shared genes was done using InterPro v 5.14-53.0 for Linux using a pipeline developed in Laboratory Evolutionary Innovation in CABD (Centro Andaluz de Biología del Desarrollo). Gene onthology terms (GO terms) were also assigned to the proteins found to be common between LF1, UC8 and FC18 identified by Prokka using Blast2GO Basic v 3.1.3 (Conesa et al., 2005).

2.10 Contigs realignment

To realign the contigs, CONTIGuator 2 (Galardini et al., 2011) was used. This is a finishing tool for structural insights of draft genomes using Artemis Comparion Tool (ACT). This tool is based in mapping the genome against the reference genome – *R. baltica* and *B. marina* in this case, using blastn algortithm with an E-value of 1e-5. In order to confirm the results obtained from CONTIGuator, several approaches were

done to assess the accuracy and validity of the result. To do this valiation, ABACAS, Mauve, MUMmer and PROmer were used.

2.11 Prophage sequences detection and genome viewer

The detection of prophage sequences within the three strains was performed with PHAST (Zhou et al., 2011). PHAST is a web server that performs a number of database comparisons to detect the presence of phage sequences. It also provides a hint at the variation in G+C content that can be related to alien DNA. Also PHAST provides information on transposons presence, not only the prophages, since the presence of a different composition in GC can be related also to Lateral Gene Transfer. CCT software allowed the analysis of the features of the three genomes under study (Grant et al., 2012). The comparisons are conducted using BLAST. The results are presented in the form of graphical maps that can also show sequence features, gene and protein names, COG category assignments, and sequence composition characteristics.

2.12 Genome mining

The prediction of antibiotics and secondary metabolite candidate genes in the three genomes were analysed using antiSMASH (antibiotics & Secondary Metabolites Analysis SHell) v 3.0.4 (Weber et al., 2015) which allows an automatic genomic identification and analysis of biosynthetic gene clusters.

3. Results and Discussion

3.1 16S rRNA gene identification and gDNA quantification

In order to confirm *Roseimaritima ulave* strain UC8, *Rubripirellula obstinata* strain LF1 and strain FC18 species identity and culture axenity, the 16S rRNA gene was sequenced and analysed. Amplification of the 16S rRNA gene was confirmed after gel electrophoresis by the presence of 1,500 bp bands (Fig. 3.1). Confirmation of the obtainment of the required amount for the sequencing of the gDNA was also visualized after gel electrophoresis (Fig. 3.1). It was observed that the size of the genomic DNA from the three strains is higher than 10,000 bp and that the gDNA from the three strains is coincident at the same level. The only visible difference is the quantity of gDNA in each strain that was higher in UC8.

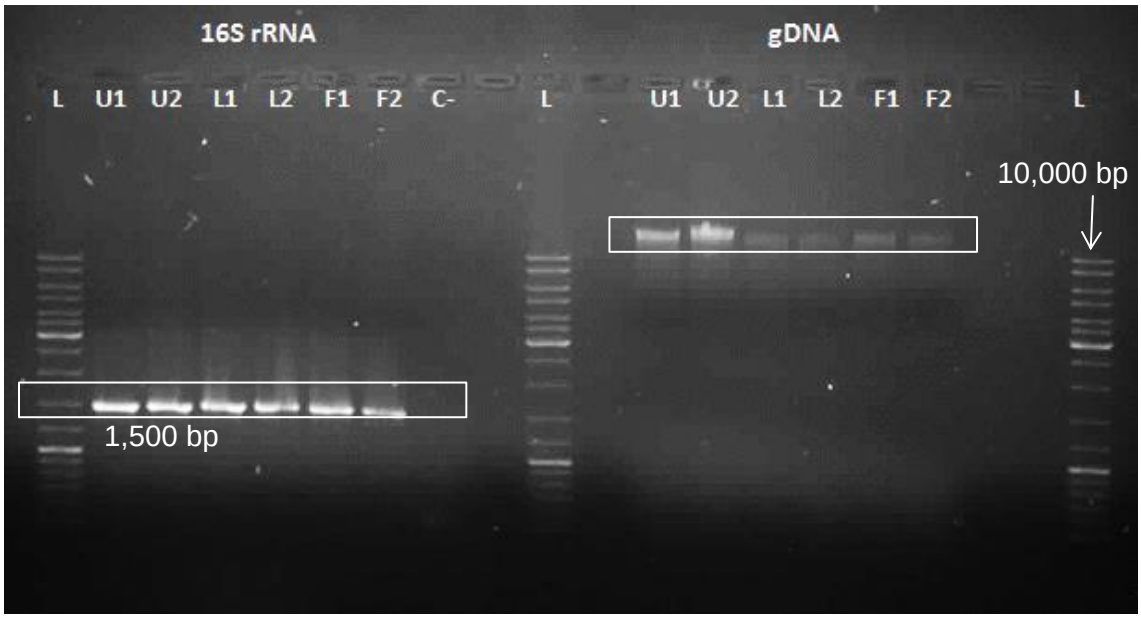


Fig. 3.1 – Gel Electrophoresis identifying the 16S rRNA gene presence and the gDNA of UC8 (U1+U2), LF1 (L1+L2) and FC18 (F1 + F2), C- - negative control, L – ladder used.

3.2 General overview of the bacterial genomes

After the assemblies, quality check and automatic annotations, it was possible to have a general overview of the bacterial genomes. In Table 3.1 are presented the general characteristics from the genomes and Table 3.2 illustrates their quality check. The

genome size of the three strains is quite big for a bacterium. Nevertheless, values between 6.6 Mbp and 8.1 Mbp are in the range of previous observations for genomes belonging to the Planctomycetaceae family (Guo et al., 2014).

When the final assemblies were uploaded to RAST, some overlap warnings in the genomes were indicated. Therefore, the contigs of each strain, only with SPAdes assembly output were *de novo* assembled in Sequencher 5.3. SPAdes is recommended for Illumina data (Bankevich et al., 2012) which supports the usage of the SPAdes output for the automatic annotation in RAST. The refined assembly did not contain duplications according to RAST. In a second step, MetaWatt had to be used to reduce the amount of possible contaminations, previously assessed by CheckM. The level of completeness of the genomes is 99.93 % for LF1, 98.77 % for UC8 and 98.77 % for FC18 with a 1.16 %, 0% and 0.11 % of contamination respectively.

Table 3.1 - General overview of the genome features from strains LF1, FC18 and LF1.

	Strains		
Attribute	**LF1**	**UC8**	**FC18**
Genome size (bp)	6,588,559	8,130,296	6,539,195
Completeness level [1]	99.93 %	98.77 %	98.77 %
Contamination	1,16%	0%	0,11%
DNA G + C content (%) [1]	54.10	59.12	53.40
CDS - RAST (bp)	5,913	5,943	5,894
CDS - Prokka (bp)	3,958	4,479	3,543
RNA genes (tRNA)*- RAST	59 (56)	64 (61)	60 (58)
tRNA genes - Prokka	69	71	66
Contigs	309	108	64
CRISPR gene [2]	1	0	1
N50 [3]	45,365	127,469	268,473
ORFs [4]	5,200	5,769	5,096

Information according to [1] CheckM; [2] RAST and Prokka ; [3] QUAST; [4] Prokka (Prodigal). * - number of RNA genes and tRNAs genes between parenthesis.

As it is acknowledged, there is potential to improve the assemblies, as an assembly has always room from improvement. However, these genomes were at the end reliable starting points to start deeper, accurate analysis and comparisons. Initially the main concern was to remove the contaminations in LF1 (9.74%) without losing the gene

content. In LF1 genome, only contigs above 1000 bp (Table 3.2) were present. This happens because after assessing the initial contamination level of 9.74%, there were paired reads not assembled to the rest that hence, were removed. Removing the contigs under 1,000 bp reduced the contamination from 9.74% to 1.16%, being nevertheless, LF1 the strain with higher level of contamination. This may have happened in the sample preparation for the sequencing step, as it was proved that the colonies were pure in before sending to sequence, by the 16S rRNA analysis. These contaminations might have altered the genome size as the total assembly size may increase, and even exceed genome size, due to contaminants (Chitsaz et al., 2011) that can contribute to multiple contigs. After this contamination correction confirmed by CheckM treatment contigs were analysed by MetaWatt but no improvement was achieved and, consequently, this directed modification of the assembly was not maintained in the assemblies.

Table 3.2 – Quality assessment of the final assembly of the three strains performed by QUAST 2.2.

Attribute	LF1	UC8	FC18
# contigs (>= 0 bp)	309	108	64
# contigs (>= 1000 bp)	309	94	60
Total length (>= 0 bp)	6,588,559	8,130,296	6,539,195
Total length (>= 1000 bp)	6,588,559	8,122,713	6,531,029
# contigs	309	102	58
Largest contig	249,803	369,108	706,834
Total length	6,588,559	8,128,167	6,536,850
N50	45,365	127,469	268,453
N75	22,205	79,236	189,232
L50	42	20	9
L75	93	40	16
# N's per 100 kbp	0.00	0.00	0.00

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (> = 0 bp)" and "Total length (>= 0 bp)" include all contigs).

In order to assess the quality of the assemblies after the *de novo* sequencing, Quast software was used. In Table 3.2 are presented the characteristics of each assembly. C+G values vary a lot in the bacterial communities, ranging between 15 % and 85 %, depending greatly on their habitats – the most complex habitats normally are associated bacteria with larger genomes, containing greater quantity of GC% (Land et al., 2015). Based on previous permanent genome draft analysis *Rhodopirellula* spp.

have around 54 % of G+C (Frank et al., 2013; Klindworth et al., 2014; Richter et al., 2014; Richter-Heitmann et al., 2014; Wegner et al., 2014). Besides, within the *Planctomycetaceae* family the G+C % varies from 50 to 67 % (Guo et al., 2014). Hence, the values presented from the three strains of 54.1% for LF1, 59.12% for UC8 and 53.40% for FC18 (Table 3.1), corroborate the previous observations. These values are also in accordance with the values referred for *Rubripirellula obstinata* LF1 and *Roseimaritima ulvae* UC8 (Bondoso et al., 2015).

Analysing the N50, the major quality parameter in the analysis of the assemblies, LF1 strain is the one showing a poorer quality (Table 3.2). Its N50 value is low when compared to the other two assemblies from FC18 and UC8 which show a satisfactory assembly quality.

3.2.1 Gene prediction

This step is the very first step one happening in the genomic annotation, a process to discover gene encoding regions. In fact, sequencing ORFs is easily detectable. Prodigal (in the Prokka pipeline). In the genomes of the three planctomycetes, 5,913 ORFs for LF1, 5,943 for UC8 and 5894 from FC18 (Table 3.1) were identified. In comparison with the closest annotated organisms, *R. baltica* SH1^T which has 7,325 putative protein encoding ORFs (Hieu et al., 2008) and *B. marina* DSM 3645 with 6,025 (Fuchsman and Rocap, 2006). This ORF number is always higher than the number of CDS from Prokka (Table. 3.1) as the annotation lowers the number of coding sequences comparing to the ORFs.

3.2.2 Gene Annotation

The genomes annotation for UC8, LF1 and FC18 was done with RAST, an automated web service, and with Prokka annotator. There are some minor differences between the results from these two annotators as Table. 3.1 shows, especially in the number of coding sequences and RNAs. This happens due to differences in gene prediction, the threshold value used, algorithms as well as the type of clustering each one of the assemblers uses. However it is always good to compare the output of both in order to have more accurate and significant results. Both of the annotators rely on BLAST search over different databases of public genomes, being therefore able to access the gene functions and metabolic roles, among others. Besides, in the end it was decided

to use the number of CDS obtained through Prokka for the differential analysis with OrthoMCL since the CDS number is more conservative, Prokka is tailored for prokaryotic genomes. Furthermore, Prokka not only uses BLAST but HMMER, which is based on HMM-HMM comparison, which is more informative.

3.3. Genome comparative analysis with RAST and SEED - viewer

After uploading the genomes and having the annotation in RAST done, it was possible to analyse the features of the genomes and do some comparative analysis using the SEED – viewer. The genes belonging to the genomes are split into subsystem categories and displayed in a pie chart, illustrated in Figure 3.1. The pie chart is labelled with the different subsystems of a typical organism, where it is possible to maximize the categories to explore specific subsections.

The contigs were annotated in RAST pipeline predicting 5,913, 5,943 and 5,894 protein encoding genes (PEGs), or CDSs, for LF1, UC8 and FC18 respectively (Table 3.1). The PEGs were classified by their metabolic function and compared to classified PEGs found in the closely related Planctomycetes' species, which have similar sized genomes. According to the RAST analysis, LF1 had 25% subsystem coverage of all known metabolic processes with 1,448 PEGs; UC8 had 29% subsystem coverage of all known metabolic processes with 1,683 PEGs and FC18 had 25% subsystem coverage with 1,452 PEGs (Table 3.3). These values compared to the 23% for *R. baltica* with 1,496 PEGs and to the 25% of *B. marina* with 1,368 PEGs, indicate that despite having higher values than the references, the three strains and also the two references still have more genes to be identified. Comparison of the genes found in the metabolic subsystems categories was overall similar with exception of the "Photosynthesis" where LF1 presented 10 genes and the "Phages, Prophages, Transposable elements, Plasmids" where LF1, UC8 and FC18 showed to have genes. "Iron acquisition subsystem" group was also present in UC8 but not in any of the others (Table 3.4; Fig 3.2). More insights on the particularities found in the annotations of LF1, UC8 and FC18 are discussed on section 3.5.

Table 3.3 – Subsystem coverage of FC18, UC8 and LF1 in the RAST annotation. *R. baltica* and *B. marina* annotation is from the database and are used as comparison organisms.

Subsystem coverage	FC18	UC8	LF1	*R.baltica*	*B. marina*
in subsystem (%)	25%	29%	25%	23%	25%
non-hypothetical	1,394	1,598	1,379	1,427	1,282
hypothetical	58	85	69	69	86
not in subsystem (%)	75%	71%	75%	77%	75%
non-hypothetical	1,231	1,397	1,141	1,428	1,327
Hypothethical	3,211	2,863	3,324	3,669	2,932

CRISPR "clustered regularly interspaced short palindromic repeats" PEGs were detected in the genome of LF1 and FC18 (Table 3.1). These genes were also identified in several planctomycetes like *Gemmata obscuriglobus* UQM 2246 (10 genes), "*Candidatus* Kuenenia stuttgartiensis" (4 genes) and *Blastopirellula marina* SH 106^T, DSM 3645 (2 genes) and absent in *Rhodopirellula baltica* SH1^T (Fuerst, 2013). These sequences play a fundamental role in the immune system of bacteria (Rath et al., 2015) and are now considered the newest tool to perform genetical engineering. So far Planctomycetes rely on the Tn5 transposase mechanism to generate mutants (Jogler et al., 2011). This new CRISPR mechanism can be a promising tool to help advancing the understanding of the planctomycetal cell biology.

Table 3.4 – Number of genes presented in the subsystem categories presented in FC18, UC8 and LF1 in RAST. *R. baltica* and *B. marina* data belongs to the database and are used as comparison organisms.

	Number of genes				
	FC18	UC8	LF1	*R. baltica*	*B.marina*
Number of subsystems	368	388	372	359	323
Subsystem Category					
Cofactors, Vitamins, Prosthetic Groups, Pigments	193	206	174	186	203
Cell Wall and Capsule	89	120	94	76	93
Virulence, Disease and Defense	103	111	79	74	70
Potassium metabolism	16	21	20	14	25
Photosynthesis	0	0	10	0	0
Miscellaneous	49	51	42	43	16
Phages, Prophages, Transposable elements, Plasmids	6	8	13	0	0
Membrane Transport	56	94	53	69	61
Iron acquisition and metabolism	0	1	0	0	0
RNA Metabolism	167	161	167	144	180
Nucleosides and Nucleotides	80	74	83	79	89
Protein Metabolism	255	267	248	207	197
Cell Division and Cell Cycle	12	18	27	22	25
Motility and Chemotaxis	79	72	74	65	90
Regulation and Cell signaling	27	15	34	31	17
Secondary Metabolism	10	10	9	9	9

	Number of genes				
	FC18	**UC8**	**LF1**	***R. baltica***	***B.marina***
DNA Metabolism	83	112	142	107	85
Fatty Acids, Lipids, and Isoprenoids	106	122	121	120	131
Nitrogen Metabolism	14	25	26	29	25
Dormancy and Sporulation	5	5	5	4	4
Respiration	50	63	56	61	95
Stress Response	97	163	69	152	59
Metabolism of Aromatic Compounds	11	7	3	5	1
Amino Acids and Derivatives	252	293	273	257	257
Sulfur Metabolism	23	47	35	46	24
Phosphorus Metabolism	61	75	44	43	52
Carbohydrates	268	280	268	241	210

3.3.1 Function based comparison

The function based comparison tool is defined by having genes for all the functional roles that compose a variant of a subsystem, defined in RAST, enabling to analyse unique functions found in any of the genomes. Table 3.5 shows the number of common functioning parts between two strains and what is singular of each one as well.

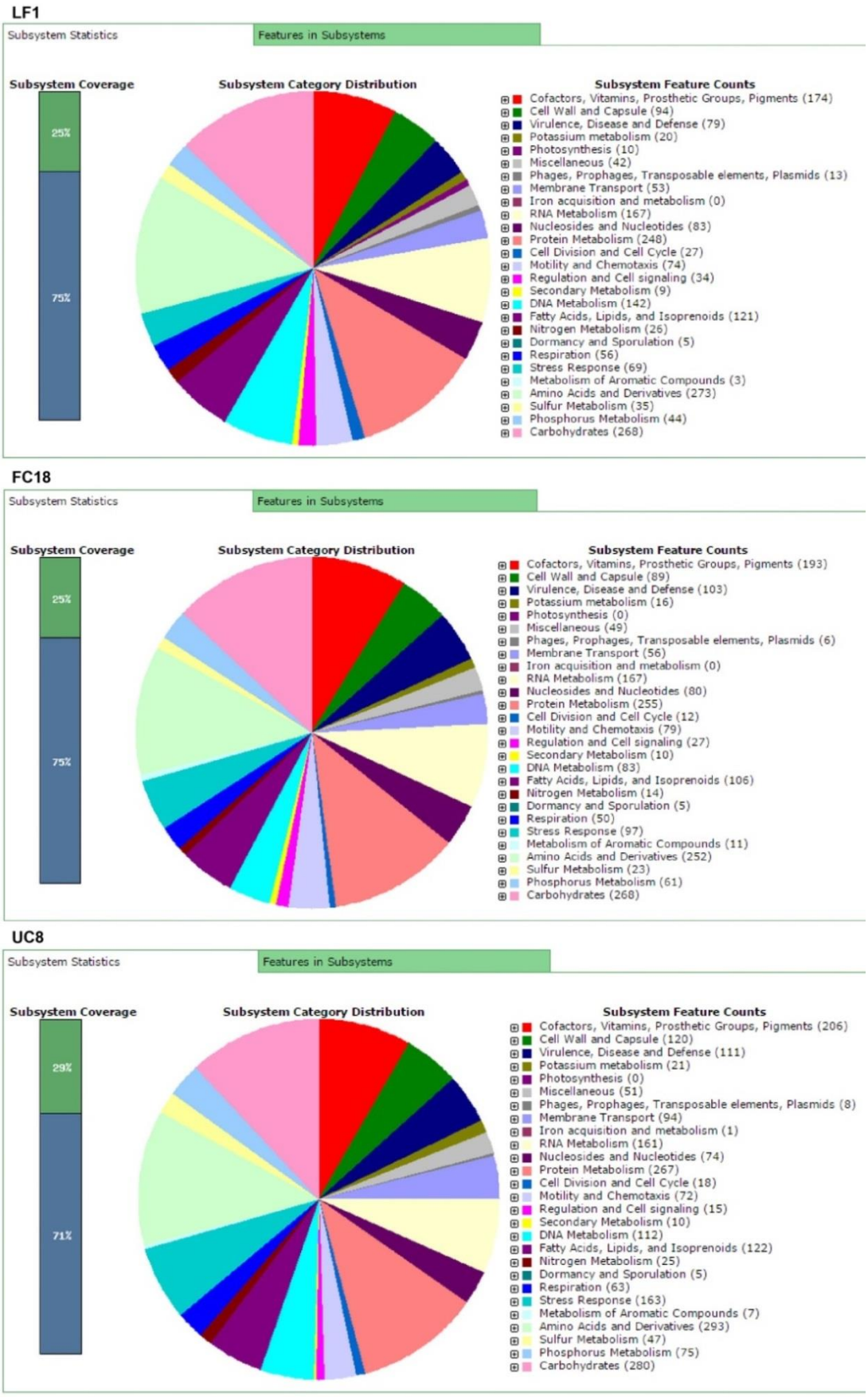


Fig. 3.2 – Pie chart showing the RAST subsystems to which each genome is connected.

Table 3.5 – Number of common and unique functioning parts of the genomes between A (reference genome) and B (comparison genome)

	Total	**A + B**		**A**		**B**	
	# of funct. parts	# of funct. parts	%	# of funct. parts	%	# of funct. parts	%
UC8 (A) and *R.b* (B)	1,857	1,533	82.55	177	9.53	127	6.84
LF1 (A) and *R.b* (B)	1,787	1,470	82.26	167	9.35	150	8.39
LF1 (A) and UC8 (B)	1,891	1,530	80.91	164	8.67	198	10.47
FC18 (A) and *R.b* (B)	1,804	1,374	76.16	211	11.70	219	12.14
FC18 (A) and *B. m* (B)	1,688	1,188	70.38	280	16.59	220	13.03
LF1 (A) and FC18 (B)	1,839	1, 416	77.00	224	12.18	199	10.82

R. b – Rhodopirellula baltica SH1^T*; B. m – Blastopirellula marina* DSM 3645

UC8 and LF1 share approximately 82% of their functioning parts with *R. baltica* and share around 80% of the functioning parts between themselves. Between FC18 and *B. marina* only share 70%, whereas FC18 and *R. baltica,* share 76% of the functioning parts. However, considering the phylogenetic relationship based on the 16S rRNA gene (Fig. 2.2), FC18 is more closely related to *B. marina* than to *R. baltica*. This may suggest that either there was a case of lateral gene transfer (LGT) or that there is a closer relationship between FC18 and *R. baltica* functionally. Futhermore, the closest neighbour identified in RAST of FC18 is *Blastopirellula marina DSM 3645* with a score of 496, followed by *Rhodopirellula baltica* SH1^T with a score of 476, corroborating the phylogenetic distance obtained. For the other two strains the closest neighbour is for both *R. baltica* SH1^T with a score of 526 and 517 for FL1 and UC8 respectively.

3.3.2 Sequence based comparison

Figure 3.3 illustrates in a circular map the sequence based comparison of the contigs/genes of the three planctomycetes strains against the reference *R. baltica*. Table 3.6 presents the comparison of the aminoacid identity of the five strains (LF1, UC8, FC18, *R. baltica* and *B. marina*) against UC8, LF1 and FC18.

Table 3.6. - Number of common sequences and bidirectional best hit in percentage. This analysis was performed in RAST.

	LF1		UC8		FC18	
	Common CDS (%)	bidirectional hit (%)	Common CDS (%)	bidirectional hit (%)	Common CDS (%)	bidirectional hit (%)
LF1						
UC8	66.4	70.3				
FC18	51.2	74.3	53.8	78.9		
***R. baltica* SH1^T**	57.3	76.2	60.5	83.7	52.0	69.7
***B. marina* DSM 3645**	57.6	66.7	66.2	71.5	49.2	54.5

Table 3.6 shows that the percentage of common CDS varied between 49.2 % and 66.*4* % and that the higher values were obtained between UC8 and LF1 followed by *B. marina* / *R. baltica* and UC8. FC18 the strain with the lowest CDS comparative values, again, showed more common protein sequence similarity with *R. baltica* than *B. marina*. Figure 3.3 evidences the low homology between the three strains and *R. baltica*, with very few CDS similarity above 80 %. BLAST Dot Plot presented in RAST indicates that there is low similarity in the genome organization as well (data not shown).

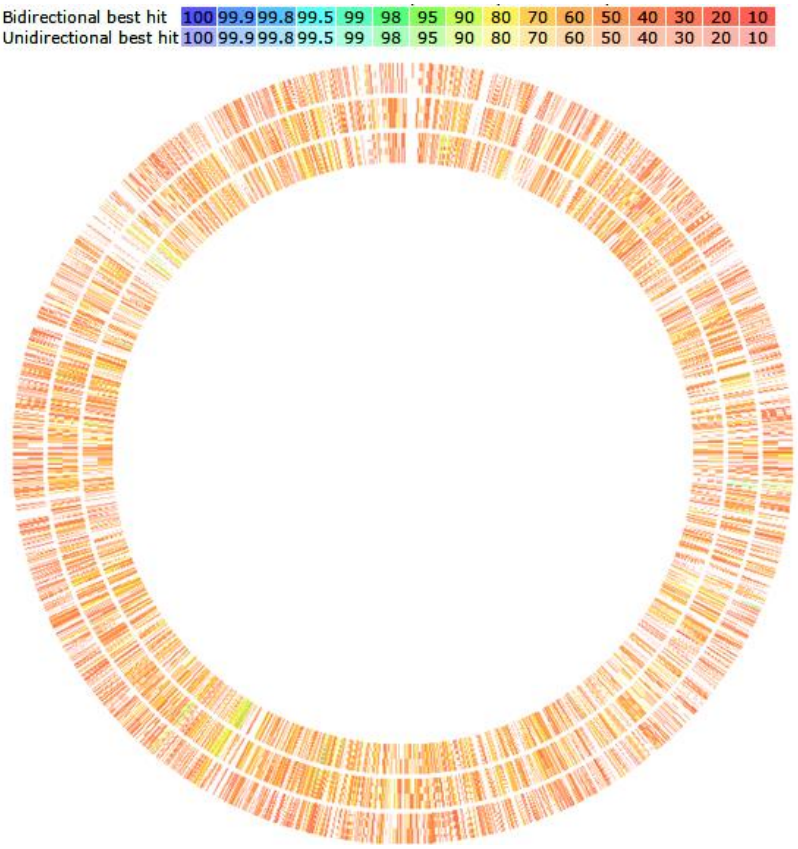


Fig. 3.3 – Circle plot showing the comparison LF1, UC8 and FC18 genomes relative to *Rhodopirellula baltica* SH1^T as reference genome (out to inside). In the legend the percent protein sequence identity is shown; the blue colour represents the highest protein sequence similarity and red represents the lowest.

3.4 Genome comparative analysis based orthologue proteins

The detection of orthologue and paralogue proteins conserved among LF1, UC8 and FC18 and between these and *R. baltica*, *B. marina* and *P. limnophilus* was done after the annotation and gene prediction with Prokka. This analysis and comparison also allowed, by exclusion, to give an overview of genes/proteins non-shared with other bacteria. The clustered proteins have high levels of similarity among each other. The non-clustered proteins (here referred as unique proteins) are the ones that demonstrate low levels of similarity.

3.4.1 Comparison between LF1, UC8 and FC18

In order to perform the clustering analysis with OrthoMCL the proteins had to be placed in clusters, allowing 70.96% in LF1, 66.63 % in UC8 and 62.09 % in FC18 (Table 3.6). The three strains under study have a total of 6,187 proteins shared in common clusters (Fig. 3.4). The proteins not clustered and thus not shared among them are 1,510 for LF1, 1,925 for UC8 and 1,932 FC18 (Fig. 3.4 and Table 3.7). As evidenced in Table 3.7 strains LF1, UC8 and FC18 have 516, 290 and 385 paralog clustered proteins respectively. The paralog proteins of each strain are related due to replication events inside of each genome. Nevertheless they might present different functions between themselves (Li et al., 2003).

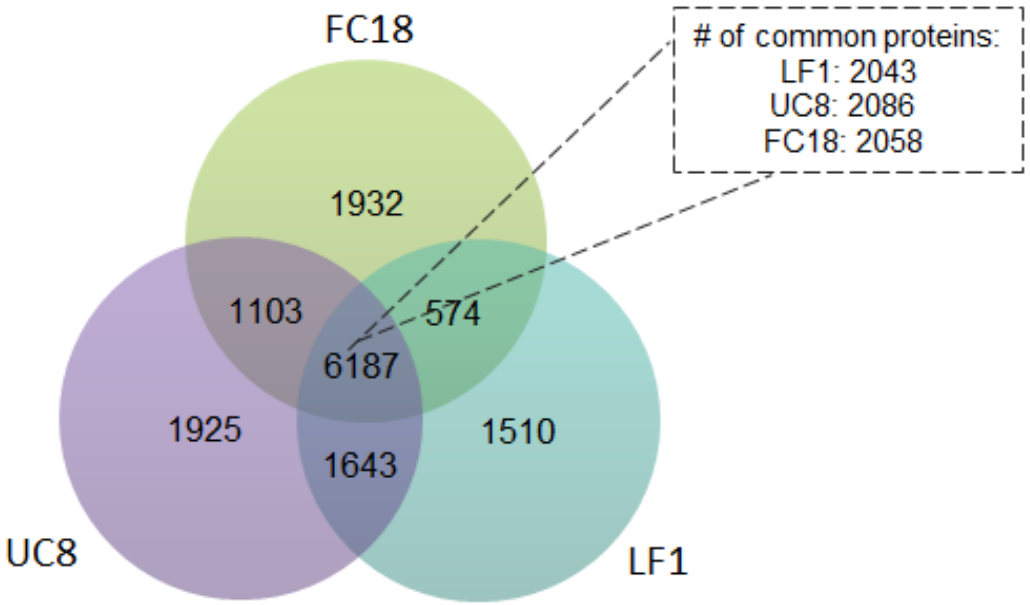


Fig. 3.4 - Number of common clustered proteins and unique proteins among LF1, UC8 and FC18.

Table 3.7 – Number of annotated CDS by Prokka annotator and the number of clustered and non-clustered (unique) CDS belonging to LF1, UC8 and FC18 assessed by OrthoMCL.

	Number of CDS/proteins		
	LF1	**UC8**	**FC18**
Total CDS annotated	5,200	5,769	5,096
Clustered (%)	3,690 (70.96%)	3,844 (66.63%)	3,164 (62.09%)
# paralogues	516	385	290
Unique (%)	1,510 (29.04%)	1,925 (33.37%)	1,932 (37.91%)

FC18 and LF1 are the ones sharing the lowest number of clustered proteins sharing approximately 280, followed by FC18 and UC8 with 541 and 562 (Table 3.8). UC8 and LF1 are the ones that share more CDS inside of the clusters, 811 and 832 respectively. These results support the 16S rRNA gene phylogenetic closeness of UC8 and LF1 and a bigger distance between these two and FC18 (Fig 2.2). In relation to the clustered orthologue proteins shared by the three strains the values are quite similar (Table 3.8).

Table 3.8 – Number of clustered orthologue proteins shared among the strains.

Comparative groups	**Number of CDS/proteins**		
	A	**B**	**C**
UC8 (A) + LF1 (B)	811	832	
UC8 (A) + FC18 (B)	562	541	
LF1 (A) + FC18 (B)	299	275	
LF1 (A) + FC18 (B) + UC8 (C)	2,043	2,058	2,086

3.4.2 Comparison between LF1, UC8, FC18, *R. baltica*, *B. marina* and *P. limnophilus*

Strains LF1, UC8 and FC18 inhabiting the macroalgae biofilm where further compared with non-macroalgal associated species, Rhodopirellula *baltica* SH1^T, *Blastopirellula marina* DSM 3,465 and *Planctomyces limnophilus* DSM 3,776. To assess if their microenvironment, the macroalgae biofilm, has any kind of influence in their genomes when compared to planctomycetes from other habitats, this differential comparison was done. The reference organisms were isolated either from the aquatic environment (first two) or from fresh water lake (the latter) and therefore might have different genes in their genome.

Table 3.9 – Number of clustered orthologue proteins shared among the strains.

Comparative groups	Number of CDS/proteins					
	A	B	C	D	E	F
UC8 (A) + LF1 (B)	127	130				
UC8 (A) + FC18 (B)	109	134				
UC8 (A) + *P. limnophilus* (B)	62	63				
UC8 (A) + *B. marina* (B)	153	160				
UC8 (A) + *R. baltica* (B)	259	264				
LF1 (A) + *B. marina* (B)	32	36				
LF1 (A) + *P. limnophilus* (B)	38	37				
LF1 (A) + *R. baltica* (B)	210	202				
FC18 (A) + LF1 (B)	110	111				
FC18 (A) + *P. limnophilus* (B)	47	46				
FC18 (A) + *B. marina* (B)	124	126				
FC18 (A) + *R. baltica* (B)	75	79				
LF1 (A) + UC8 (B) + FC18 (C)	39	46	43			
LF1 (A) + UC8 (B) + FC18 (C) + *B. marina* (D) + *R. baltica* (E) + *P. limnophilus* (F)	1,278	1,325	1,307	1,309	1,299	1,278

When comparing the six planctomycetes all of them present an approximate number of proteins in their shared cluster groups (Table 3.9). Also, there are 128 (39 + 46 +43) clustered proteins only shared by LF1, UC8 and FC18 and that are the main focus of this analysis. In Appendix I the table shows the clusters and respective clustered proteins belonging to this comparative group (LF1 + UC8 + FC18). Insights on these particular proteins will be given in section 3.6.

The shared genes among LF1, UC8 and FC18 were also assigned to Gene Ontology (GO) terms. The GO types separated the proteins into molecular function, biological process and cellular component groups (Fig. 3.5). The majority of terms used in LF1, UC8 and FC18 annotations are related to the biological processes and molecular functions with 56 and 72 CDSs assigned to each type, respectively. Cellular components had 20 CDSs assigned to the group. It is important to highlight that in the GO terms each protein can be placed in more than one group at a time (Fig. 3.5, Appendix III). GO terms divide the CDS in ontology levels, meaning that there are terms inside of terms following a specific hierarchical structure of the Gene Ontology (Appendix III).

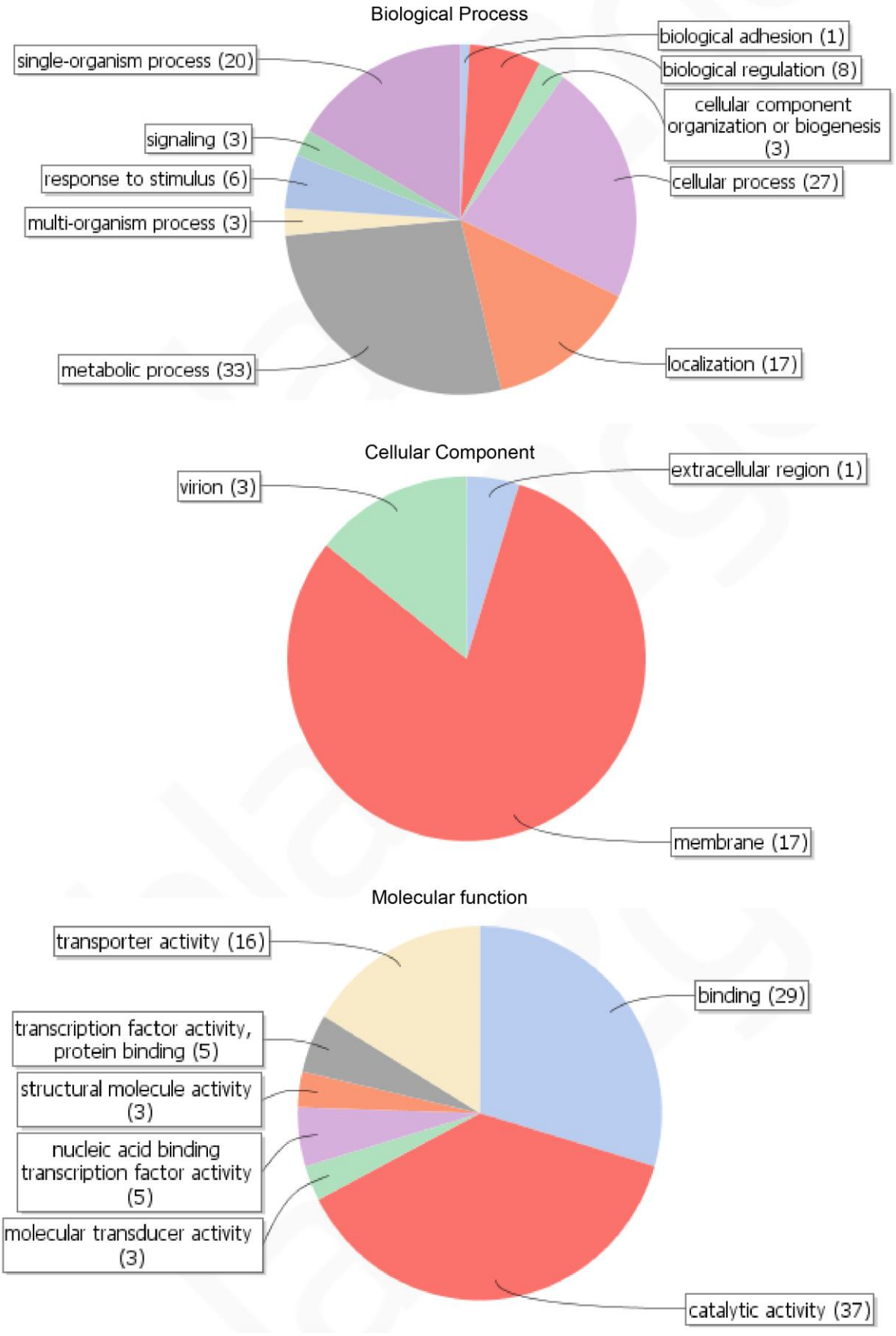


Fig. 3.5 – GO terms (belonging to level two) mapped in the common genes among LF1, UC8 and FC18 retrieved with blast2GO.

3.5. Further characterisation of the bacterial genomes

In the characterization of the bacterial genomes, phage sequences were detected using specific software (PHAST). These analyses showed that all the planctomycetes have phage sequences. In *Rubripirellula obstinata* strain LF1 there is one prophage region with 17 CDSs placed in the region between 7.1 Mbp and 8.6 Mbp of the genome with a length of 14.4 Kbp (Fig. 3.6). In *Roseimaritima ulvae strain* UC8 there are 9 CDSs with a length of 8.9 Kbp. Also, in strain FC18 there is one prophage region as well, with 28 CDSs from 3.9 Mbp to 6.1 Mbp with a region length of 21.9 Kbp. Despite all of the general presence of phage sequences, many are hypothetical and phage-like proteins. LF1 and FC18 displayed an attachment site CDS where it can be possible the integration of the bacteriophage in the host genomes. FC18 was also shown to have coding sequences for transposases (Fig. 3.6). The presence of phage-like sequences in planctomycetes and knowing the phage role as gene vehicles could also support the exchange of genes responsible for antibiotic resistances and others, between different strains and even species that can happen as well in Planctomycetes (Mazaheri Nezhad Fard et al., 2011). Combining these results with the genome circular overview (performed with CCT) (Fig. 3.7) allowed to deeply understand the genomic organisation of the planctomycetes strains. Looking at the genomes on Figure 3.7 from outside to inside, it is possible to see the CDS identified and clustered within a specific COG group (Fig. 3.7); on the next track it is shown the alignment with *Rhodopirellula baltica* $SH1^T$ which is around the 94% identity, overall. The G+C content distribution is greatly distributed. However, it can be seen some areas with a higher distribution in one strand, something abnormal that can be connected with the presence of phage sequences, cases of LGT or transposons. These observations can be supported by with the aforementioned PHAST analysis results showed. In the last track, the GC skew is presented and, in both LF1 and UC8, it is very complicated to identify the replication origin (ORI) and terminus.

Previous genomic studies performed with *R. baltica* (Glöckner et al., 2003)*, Pirellula staleyi* (Clum et al., 2009) show a balanced GC skew with an easy detectable ORI and terminus. This may evidence a poor assembly process; the other option for this result could have been a bad sequencing due to a low sequencing coverage. However the coverage was high (around 50x) being, thus, most likely the problem to be connected with the assembly process. In fact, in order to realign the contigs from LF1, UC8 and FC18 the CONTIGuator software detected some problems.

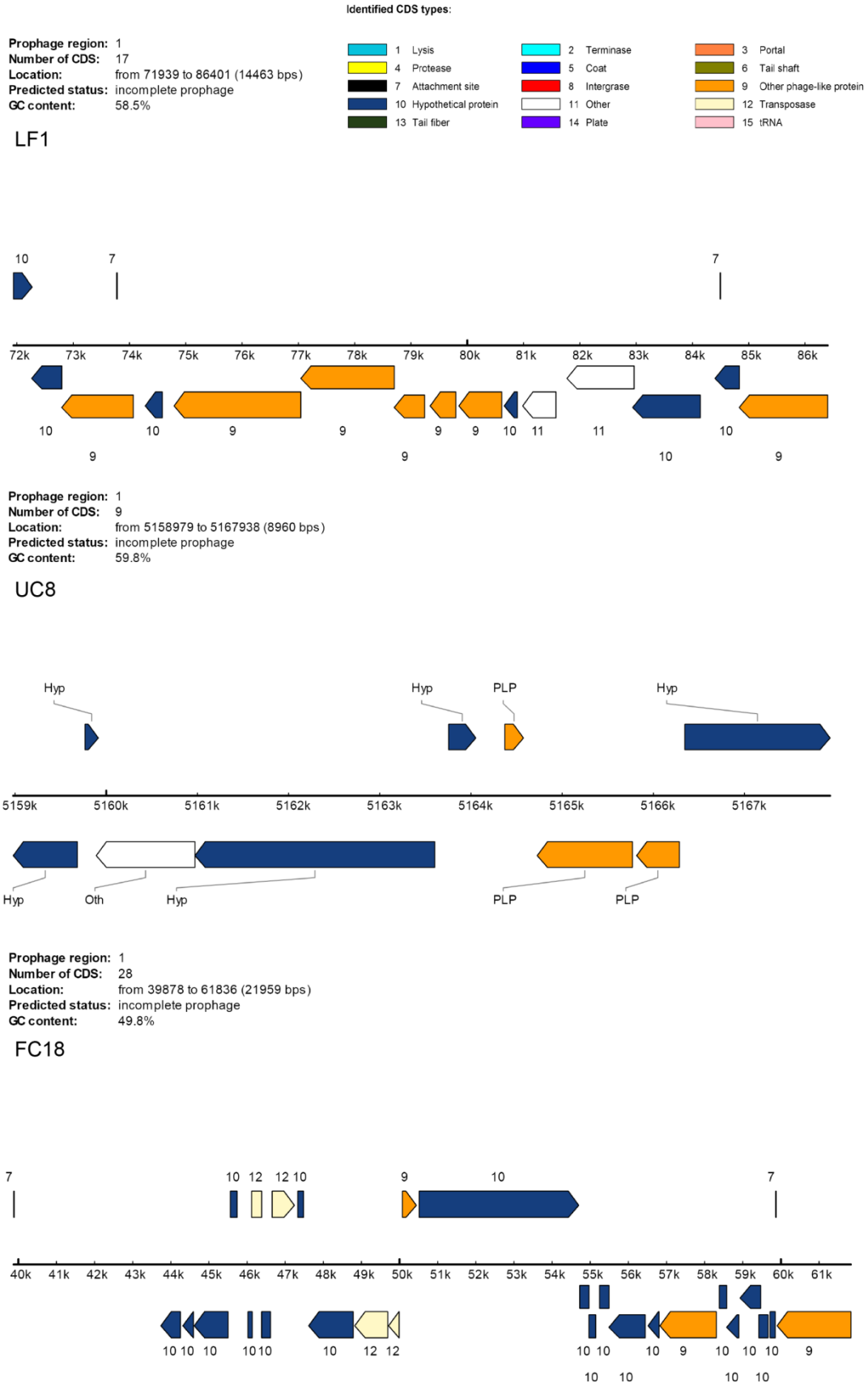


Fig. 3.6 – CDS in LF1, UC8 and FC18 retrieved by PHAST

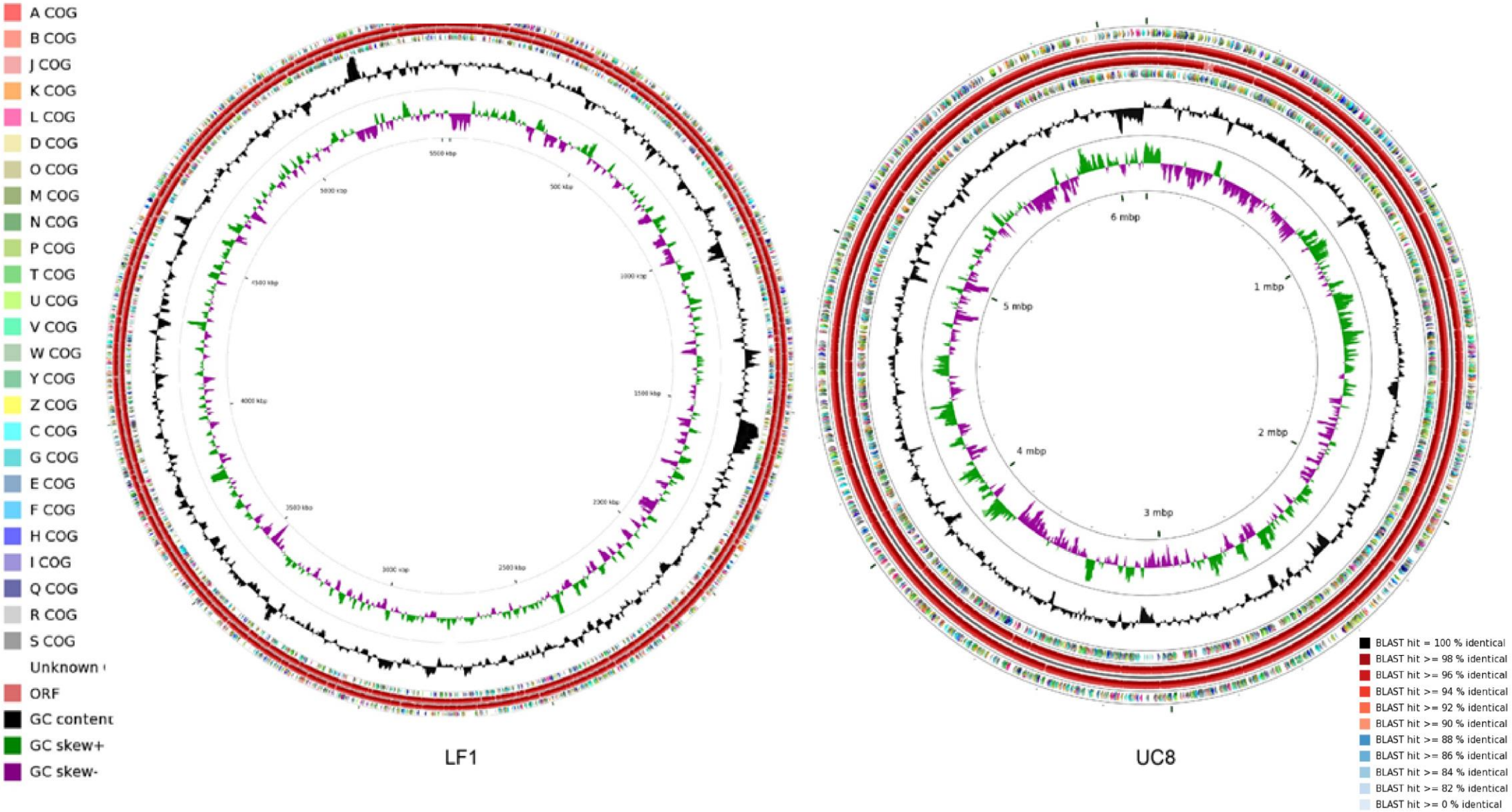


Fig. 3.7 – Circular view of the genome from LF1 and UC8 obtained from CCT software. Legend presented on the left shows the COG groups, ORFs and GC content and skew; on the right the BLAST hits against *R. baltica* SH1^{T}.

Firstly, FC18 genome, more phylogenetically related with *B. marina* by the 16S rRNA gene, could not be mapped against it. Therefore, it was mapped against *R. baltica* being, nevertheless, only 8 contigs out of 64 mapped and losing more than 1.0 Mbp (Appendix II). As a result CCT software was not able to perform FC18 analysis and no COGs or circular genome image were retrieved. Both LF1 and UC8 also did not map their contigs entirely against *R. baltica* as well. They also lost approximately 1 Mbp after the realignment with CONTIGuator (Appendix III). In the three cases, great parts of the unmapped contigs were labelled as poor coverage contigs, supporting therefore the possible problems in the assembly process.

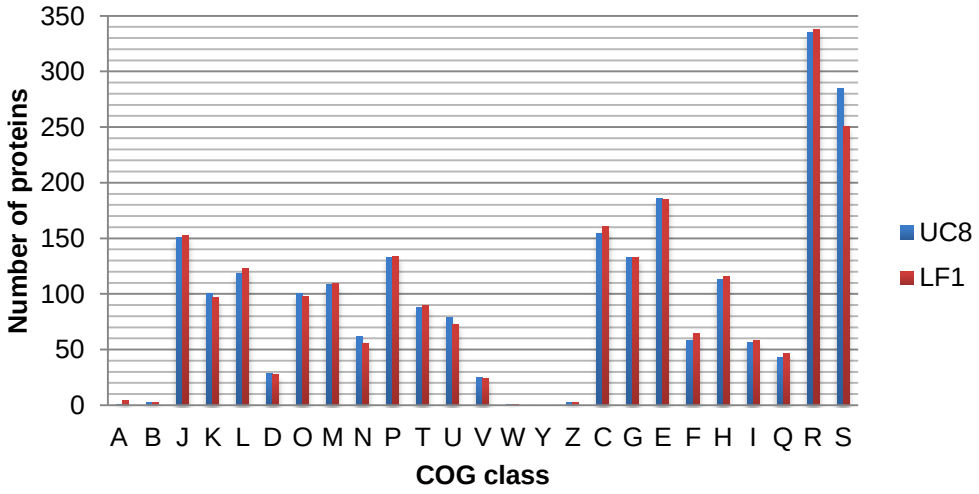


Fig. 3.8 – COG classes distribution of LF1 and UC8, data retrieved by CCT.

For the COG classes distribution in LF1 and FC18 is important to highlight the loss of some CDS and groups, due to the realignment of the contigs and respective deletion of the unmapped ones. In any case, it is still possible to observe that both LF1 and FC18 share a similar amount of orthologues in each group (Fig. 3.8), being the COG groups of “Amino acid transport and metabolism”, “Translation, ribosomal structure and biogenesis”, “Energy production and conversion”, “Carbohydrate transport and metabolism” and “Inorganic Ion transport and metabolism” the ones with higher quantity of orthologue proteins (Table 3.10). Many of them belonged to the “Function unknown” and “General function prediction only” showing a wide large amount of proteins needed to be characterised in these strains.

Table 3.10 - GOG classes description and number of distributed genes.

COG Class	Description	UC8	LF1
A	RNA processing and modification	1	4
B	Chromatin structure and dynamics	3	3
J	Translation, ribosomal structure and biogenesis	151	153
K	Transcription	101	97
L	Replication, recombination and repair	119	123
D	Cell cycle control, cell division, chromosome partitioning	29	28
O	Post-translational modification, protein turnover, and chaperones	101	98
M	Cell wall/membrane/envelope biogenesis	109	110
N	Cell motility	62	56
P	Inorganic ion transport and metabolism	133	134
T	Signal transduction mechanisms	88	90
U	Intracellular trafficking, secretion, and vesicular transport	79	73
V	Defense mechanisms	25	24
W	Extracellular structures (this doesn't appear in reference database)	1	1
Y	Nuclear structure (this appears once in reference database)	0	0
Z	Cytoskeleton	3	3
C	Energy production and conversion	155	161
G	Carbohydrate transport and metabolism	133	133
E	Amino acid transport and metabolism	186	185
F	Nucleotide transport and metabolism	58	65
H	Coenzyme transport and metabolism	113	116
I	Lipid transport and metabolism	57	58
Q	Secondary metabolites biosynthesis, transport, and catabolism	43	47
R	General function prediction only (examples include "Predicted thioesterase", "Predicted ATPase")	335	338
S	Function unknown (examples include "Uncharacterised conserved protein", "Predicted small secreted protein")	285	251

3.6 Shared genome features of LF1, UC8 and FC18

Annotation outputs from RAST, Prokka and Blast2GO were analysed and compared with the objective of finding particularities among these biofilm attached bacteria. Bellow, are presented curious features shared among LF1, UC8 and FC18.

Phage sequences: In the three strains annotations some capsid proteins from bacteriophage were detected. After BLAST in Uniprot and a sequence identity of 100% and an e-value of 0.0 these sequences were connected to Enterobacteria phage (Phage phi-X174), other cluster of proteins were related in homology with a protein ea22 belonging to the Enterobacteria phage lambda and a bacteriophage scaffolding protein (Prot 3629 and 3647) (Appenix I). This was also possible to be evidenced with PHAST result where some phage-like proteins were detected. Up to date, phage Pi-89 was found in *P. limnophilus* and Pi-57 was found infecting *Pirellula staleyi* (Fuerst, 2013). RAST and Blast2Go also detected some viroid and phage capsid proteins supporting even more its presence in all of the three strains (Appendix II and IV).

Phosphate metabolism: In RAST annotation table (Appendix IV) it is easily detectable among LF1, UC8 and FC18 a detection of PEGs related with phosphate metabolism, more precisely uptake and transport systems. This presence may be an ecological evidence of a competition against macroalgae for phosphate. It is still not very well know if planctomycetes are strong or weak competitors for phosphorous (Pollet et al., 2014). On the other hand, macroalgae are acknowledged to grow easily in areas with high abundance of phosphate (Kuffner and Paul, 2001).

Gluconeogenic pathway: Placed in the Pyruvate metabolism I: anaplerotic reactions, PEP, protein Phosphoenolpyruvate carboxykinase (EC 4.1.1.49) PEGs known as PEPCK were detected in the RAST automatic annotation in the three planctomycetes strains. It catalyzes metal-nucleotide coupled reversible decarboxylation and phosphorylation between phosphoenolpyruvate (PEP) and oxaloacetate (OAA) depending on the system and the availability of the intermediate. PEPCK was found in *Candidatus* Kuenenia stuttgartiensis (Aich and Delbaere, 2007) and is now known to be present in thirteen planctomycetes genomes (Flores et al., 2014). Furthermore, when this sequence present in the three planctomycetes was blasted it got an e-value of 0.0 to *Rhodopirellula maiorca* SM1 (80% identity), *Isosphaera pallida* ATCC 43644 (70% identity) and *Rhodotermus marinus* (67 % identity). Two of of them isolated from marine environments and *I. pallida* isolated from an algal mat in fresh water hot spring (Göker et al., 2011). PEPCK is also involved in the pathway of the biosynthesis of secondary metabolites and antiobitics, suggesting a possible relation of these strains to a potential secondary metabolic activity, which is very important in bacteria living in

complex biofilm environments like those on macroalgae surfaces (Lage and Bondoso, 2014).

Chemotaxy: RAST annotated a motility and chemotaxis category to one of the proteins among LF1, UC8 and FC18. In fact, chemotaxis has been connected with the transcriptomes of macroalgae-associated microbiome (de Oliveira et al., 2012). Some PEGs were connected to signal transduction histidine kinase CheA protein, which place an important role in the cellular adaptation to environmental conditions and stresses (Dutta et al., 1999) (Appendix IV). This domain was reported to be fundamental in the recognition of the surface of macroalgae in the biofilm formation (de Oliveira et al., 2012)

Stress Response: A cluster of chaperone homologue proteins, prot 3592, were detected and related with the chaperone protein DnaJ (Appendix I). This protein acts in the stress response. In this case, its presence can allow the bacteria to cope better with the oxygen toxicity when attached to the algal surface, as the levels are high. Chaperone protein ClpB (cluster Prot 1497, Appendix IV) is also related with DnaJ (Fredriksson et al., 2005; de Oliveira et al., 2012). Besides, they are also related with heat-shock proteins, commonly found in the shared clusters Prot 1497, 2870, for example (Appendix I). Moreover, the presence of PEGs encoding for Lactoylglutathione lyase (EC 4.4.1.5) was evidenced by RAST annotation. This protein is also connected with redox reactions, specially related with glycolysis. As bacteria in the biofilm of macroalgae are huge consumers of organic matter (Cottrell and Kirchman, 2000) produced by the algae, they need enzymes to convert the cytotoxic metabolic by-products, such as methylglyoxal during glycolysis (Allaman et al., 2015)

Metal binding systems: Many clustered groups of proteins were detected to be connected with copper ion binding and transmembrane transport (Prot 1537, Prot 2274, Prot 2762) (Appendix I). This presence can be related with the microenvironment from where LF1, UC8 and FC18 were isolated. All of them were isolated from macroalgae in rocky pools located in areas subjected to anthropogenic pressure, namely from industrial activities and urban sewage, which are normally rich in metals. Macroalgae easily accumulate copper and other metals from the environment in their tissues (Huang et al., 2010) potentially creating conditions of high metal levels in their biofilms. Epibiotic bacteria have thus, to cope with this and develop scavenging metabolisms.

3.7 Genome mining of LF1, FC18 and FC18

For the analysis of the antibiotic and secondary metabolites gene candidates, a comprehensive genome mining approach employing the antiSMASH secondary metabolite identification was performed. A total of 24 clusters were detected, putatively related with the production of secondary metabolites within the three genomes (Table 3.11). UC8 was the strain presenting more cluster candidates, 9, related with the production of secondary metabolites. *B. marina* and *R. baltica* have been reported to have 12 and 10 clusters respectively (Jeske et al., 2013). Many of the candidate genes for the production of secondary metabolites from LF1, UC8 and FC18 are connected with polyketide synthases (PKS) and nonribosomal peptide synthethases (NRPS). These enzymes enable the production of a myriad of bioactive molecules that can show antibacterial, antiviral, antifungal and others (Donadio et al., 2007; Wagner-Döbler et al., 2002). Two type I PKS, 2 type III PKS and one NRPS-type I PKS hybrid cluster were detected in the genomes of LF1, UC8 and FC18 (Table 3.11).

Table 3.11 – Clusters detected in the genome of LF1, FC18 and FC18.

	Clusters	Candidate genes	Description
LF1	2	linaridin	famylin of the class of Lantibiotics
	2	PKS-III	chalcone and stilbene synthase (UV protection and antifungal defense)
	1	terpene	various predictions
	1	resorcinol	-
	1	PKS-I	-
UC8	2	PKS-III	chalcone and stilbene synthase (UV protection and antifungal defense)
	3	PKS-I	two involved in resistance to chloramphenicol; one related to pullulanase (enzyme involved in the production of etanol and sweeteners)
	2	NRPS	one with resistance to tetracycline
	1	linaridin	famylin of the class of Lantibiotics
	1	terpene	phythoene synthase
FC18	4	NRPS-PKSI	several different antibiotcs predicted
	3	terpene	phythoene synthase
	1	linaridin	famylin of the class of Lantibiotics

Besides PKS and NRPKs, putative encoding genes for lantibiotics were also detected in the genomes of the three strains. They are produced by Gram-positive bacteria and are shown to gave a strong antimicrobial activity against Gram-positive bacteria and nowadays, some clinical applications are in trials and are being proposed (Willey and van der Donk, 2007). Putative terpenes encoding genes were detected as well in these three strains. They can have antimicrobial activity, act as hormones or as vitamins and

are generally plant or fungal metabolites (Yamada et al., 2015). In this case, they are related with putative phytoene synthase connected to the carotenoid production having several mechanisms as UV protectors, as antioxidants, and as anti-inflammatory agents (Toledo-Ortiz et al., 2010). These observations demonstrate the relation between the genome size and the capacity for encoding secondary metabolite related genes (Jeske et al., 2013). These potential genes may be related to and an adaptive consequence of the complex microenvironment of bacteria living in the biofilm of macroalgae.

4. Conclusion and future perspectives

In general, the genomes from *Rubripirellula obstinata*, strain LF1, *Roseimaritima ulvae*, strain UC8 and the still uncharacterised strain FC18 shared a great similarity with the one of *Rhodopirellula baltica* SH1^{T,} especially in the general genome features. Despite the closer phylogenetic closeness of FC18 to *Blastopirellula marina* DSM 3645, this study shows that it share more functional groups with *R. baltica* SH1^{T}, suggesting a closer phylogenetic relationship of both strains and/or the occurrence of some eventual cases of LGT. Further studies are needed to assess the most accurate phylogenetic position of FC18 using other phylogenetic markers.

One hundred and twenty eight homologue coding sequences sharing high levels of similarity among them and not shared with *R. baltica* SH1^{T}, *P. limnophilus* and *B. marina* DSM 3645 were detected. These CDSs revealed some particular characteristics that might be connected with the complex habitat from where these three planctomycetes were isolated. These features include, for instance, chemotaxy (important in the biofilm formation), stress response (macroalgae are exposed to high levels of temperature and pollutants), carbon metabolism (macroalgae are hot spots of organic material production) and phosphate uptake (possible ecological competition against macroalgae and other microorganisms).

The several analysis performed in the genomes showed that the content of the assembly still need some improvements, allowing the realignment of all the contigs without losing possible meaningful genomic information. Therefore the localization of the error in the assembly and also an improvement in the quality of the related annotation would make many other peculiar characteristics among these three strains be evidenced. Completeness of the draft genomes could be a future goal.

Finding phage-like sequences as well as CRISPR genes in the genomes can also help develop genetic tools, just as it has helped develop genetic systems in many other bacteria, and it can be more thoroughly studied in the near future.

Other interesting future work could be focused again in the comparative genomics, this time highlighting the divergence of gene family clusters and biological processes the “unique proteins” of each strain.

References

Abed, R.M.M., Musat, N., Musat, F., and Mussmann, M. (2011). Structure of microbial communities and hydrocarbon-dependent sulfate reduction in the anoxic layer of a polluted microbial mat. Mar. Pollut. Bull. *62*, 539–546.

Aich, S., and Delbaere, L.T.J. (2007). Phylogenetic study of the evolution of PEP-carboxykinase. Evol. Bioinform. Online *3*, 333–340.

Allaman, I., Bélanger, M., and Magistretti, P.J. (2015). Methylglyoxal, the dark side of glycolysis. Front. Neurosci. *9*, 23.

Andrew, D.R., Fitak, R.R., Munguia-Vega, A., Racolta, A., Martinson, V.G., and Dontsova, K. (2012). Abiotic factors shape microbial diversity in Sonoran Desert soils. Appl. Environ. Microbiol. *78*, 7527–7537.

Armstrong, E., Yan, L., Boyd, K.G., Wright, P.C., and Burgess, J.G. (2001). The symbiotic role of marine microbes on living surfaces. Hydrobiologia *461*, 37–40.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., et al. (2008). The RAST Server: rapid annotations using subsystems technology. BMC Genomics *9*, 75.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. a, Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477.

Barion, S., Franchi, M., Gallori, E., and Di Giulio, M. (2007). The first lines of divergence in the Bacteria domain were the hyperthermophilic organisms, the Thermotogales and the Aquificales, and not the mesophilic Planctomycetales. Biosystems. *87*, 13–19.

Bengtsson, M.M., and Øvreås, L. (2010). Planctomycetes dominate biofilms on surfaces of the kelp Laminaria hyperborea. BMC Microbiol. *10*, 261.

Binnewies, T.T., Motro, Y., Hallin, P.F., Lund, O., Dunn, D., La, T., Hampson, D.J., Bellgard, M., Wassenaar, T.M., and Ussery, D.W. (2006). Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. Funct. Integr. Genomics *6*, 165–185.

Bondoso, J., Albuquerque, L., Nobre, M.F., Lobo-da-Cunha, A., Da Costa, M.S., and Lage, O.M. (2011). Aquisphaera giovannonii gen. nov., sp. nov., a planctomycete isolated from a freshwater aquarium. Int J Syst Evol Microbiol *61*, 2844–2850.

Bondoso, J., Balagué, V., Gasol, J.M., and Lage, O.M. (2014a). Community composition of the Planctomycetes associated with different macroalgae. FEMS Microbiol. Ecol. *88*, 445–456.

Bondoso, J., Albuquerque, L., Lobo-da-Cunha, A., da Costa, M.S., Harder, J., and Lage, O.M. (2014b). Rhodopirellula lusitana sp. nov. and Rhodopirellula rubra sp. nov., isolated from the surface of macroalgae. Syst. Appl. Microbiol. *37*, 157–164.

Bondoso, J., Albuquerque, L., Nobre, M.F., Lobo-da-Cunha, A., da Costa, M.S., and Lage, O.M. (2015). Roseimaritima ulvae gen. nov., sp. nov. and Rubripirellula obstinata gen. nov., sp. nov. two novel planctomycetes isolated from the epiphytic community of macroalgae. Syst. Appl. Microbiol.

Brochier, C., and Philippe, H. (2002). Phylogeny: a non-hyperthermophilic ancestor for bacteria. Nature *417*, 244.

Brown, C.T., Crusoe, M.R., Edvenson, G., Fish, J., Howe, A., McDonald, E., Nahum, J., Nanlohy, K., Ortiz-Zuazaga, H., Pell, J., et al. (2014). The khmer software package: enabling efficient sequence analysis.

Case, R.J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W.F., and Kjelleberg, S. (2007). Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl. Environ. Microbiol. *73*, 278–288.

Cayrou, C., Sambe, B., Armougom, F., Raoult, D., and Drancourt, M. (2013). Molecular diversity of the Planctomycetes in the human gut microbiota in France and Senegal. APMIS *121*, 1082–1090.

Chitsaz, H., Yee-Greenbaum, J.L., Tesler, G., Lombardo, M.-J., Dupont, C.L., Badger, J.H., Novotny, M., Rusch, D.B., Fraser, L.J., Gormley, N.A., et al. (2011). Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. Nat. Biotechnol. *29*, 915–921.

Chouari, R., Paslier, D. Le, Daegelen, P., Ginestet, P., Weissenbach, J., Sghir, A., Services, O., and Pecq, L. (2003). Molecular Evidence for Novel Planctomycete Diversity in a Municipal Wastewater Treatment Plant. *69*, 7354–7363.

Clum, A., Tindall, B.J., Sikorski, J., Ivanova, N., Mavrommatis, K., Lucas, S., Glavina, T., Del Rio, Nolan, M., Chen, F., et al. (2009). Complete genome sequence of Pirellula staleyi type strain (ATCC 27377). Stand. Genomic Sci. *1*, 308–316.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics *21*, 3674–3676.

Cottrell, M.T., and Kirchman, D.L. (2000). Natural assemblages of marine proteobacteria and members of the Cytophaga-Flavobacter cluster consuming low- and high-molecular-weight dissolved organic matter. Appl. Environ. Microbiol. *66*, 1692–1697.

Cox, M.P., Peterson, D.A., and Biggs, P.J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics *11*, 485.

D J, L. (1991). 16S/23S rRNA sequencing. Stackebrandt E, Goodfellow M, Ed. Nucleic Acid Tech. Bact. Syst. Chichester, United Kingdom John Wiley Sons 115–175.

Devoid, S., Overbeek, R., DeJongh, M., Vonstein, V., Best, A.A., and Henry, C. (2013). Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. Methods Mol. Biol. *985*, 17–45.

Devos, D.P., and Reynaud, E.G. (2010). Evolution. Intermediate steps. Science *330*, 1187–1188.

Donadio, S., Monciardini, P., and Sosio, M. (2007). Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. Nat. Prod. Rep. *24*, 1073–1109.

Dutta, R., Qin, L., and Inouye, M. (1999). Histidine kinases: diversity of domain organization. Mol. Microbiol. *34*, 633–640.

Fischer, S., Brunk, B.P., Chen, F., Gao, X., Harb, O.S., Iodice, J.B., Shanmugam, D., Roos, D.S., and Stoeckert, C.J. (2011). Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster

proteomes into new ortholog groups. Curr. Protoc. Bioinformatics *Chapter 6*, Unit 6.12.1–19.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., and Merrick, J.M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496–512.

Flores, C., Catita, J.A.M., and Lage, O.M. (2014). Assessment of planctomycetes cell viability after pollutants exposure. Antonie Van Leeuwenhoek *106*, 399–411.

Forterre, P., and Gribaldo, S. (2010). Bacteria with a eukaryotic touch: a glimpse of ancient evolution? Proc. Natl. Acad. Sci. U. S. A. *107*, 12739–12740.

Frank, C.S., Klockow, C., Richter, M., Glöckner, F.O., and Harder, J. (2013). Genetic diversity of Rhodopirellula strains. Antonie Van Leeuwenhoek *104*, 547–550.

Fredriksson, A., Ballesteros, M., Dukan, S., and Nyström, T. (2005). Defense against protein carbonylation by DnaK/DnaJ and proteases of the heat shock regulon. J. Bacteriol. *187*, 4207–4213.

Fuchsman, C.A., and Rocap, G. (2006). Whole-genome reciprocal BLAST analysis reveals that planctomycetes do not share an unusually large number of genes with Eukarya and Archaea. Appl. Environ. Microbiol. *72*, 6841–6844.

Fuchsman, C.A., Staley, J.T., Oakley, B.B., Kirkpatrick, J.B., and Murray, J.W. (2012). Free-living and aggregate-associated Planctomycetes in the Black Sea. FEMS Microbiol. Ecol. *80*, 402–416.

Fuerst, J. (2013). Planctomycetes: Cell Structure, Origins and Biology (Humana Press).

Fuerst, J. a (2005). Intracellular compartmentation in planctomycetes. Annu. Rev. Microbiol. *59*, 299–328.

Fuerst, J.A. (1995). The planctomycetes: emerging models for microbial ecology, evolution and cell biology. Microbiology *141*, 1493–1506.

Fuerst, J. a, and Sagulenko, E. (2011). Beyond the bacterium: planctomycetes challenge our concepts of microbial structure and function. Nat. Rev. Microbiol. *9*, 403–413.

Fuerst, J.A., and Sagulenko, E. (2012). Keys to eukaryality: planctomycetes and ancestral evolution of cellular complexity. Front. Microbiol. *3*, 167.

Fuerst, J. a, Gwilliam, H.G., Lindsay, M., Lichanska, a, Belcher, C., Vickers, J.E., and Hugenholtz, P. (1997). Isolation and molecular identification of planctomycete bacteria from postlarvae of the giant tiger prawn, Penaeus monodon. Appl. Environ. Microbiol. *63*, 254–262.

Fukunaga, Y., Kurahashi, M., Sakiyama, Y., Ohuchi, M., Yokota, A., and Harayama, S. (2009). Phycisphaera mikurensis gen. nov., sp. nov., isolated from a marine alga, and proposal of Phycisphaeraceae fam. nov., Phycisphaerales ord. nov. and Phycisphaerae classis nov. in the phylum Planctomycetes. J. Gen. Appl. Microbiol. *55*, 267–275.

Galardini, M., Biondi, E.G., Bazzicalupo, M., and Mengoni, A. (2011). CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. Source Code Biol. Med. *6*, 11.

Garrity, G.M., and Holt, J.G. (2001). The Road Map to the Manual. In Bergeys Manual of Systematic Bacteriology The Archaea and the Deeply Branching and Phototrophic Bacteria, D.R. Boone, R.W.

Castenholz, and George M Garrity, eds. (Springer), pp. 119–166.

Garrity, G., Rainey, F.A., and Widdel, F. (2005). Bergey's Manual of Systematic Bacteriology (Springer).

Gimesi, N. (1924). I: Planctomyces bekefii Gim nov. gen. et sp. [in Hungarian with German translation]. NHydrobiologia Tanulmanyok (Hydrobiologische Stu Dien).

Giovannoni, S.J., Godchaux, W., Schabtach, E., and Castenholz, R.W. (1987). Cell wall and lipid composition of Isosphaera pallida, a budding eubacterium from hot springs. J. Bacteriol. *169*, 2702–2707.

Giovannoni, S.J., Britschgi, T.B., Moyer, C.L., and Field, K.G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. Nature *345*, 60–63.

Di Giulio, M. (2003). The ancestor of the Bacteria domain was a hyperthermophile. J. Theor. Biol. *224*, 277–283.

Glöckner, F.O., Kube, M., Bauer, M., Teeling, H., Lombardot, T., Ludwig, W., Gade, D., Beck, a, Borzym, K., Heitmann, K., et al. (2003). Complete genome sequence of the marine planctomycete Pirellula sp. strain 1. Proc. Natl. Acad. Sci. U. S. A. *100*, 8298–8303.

Goecke, F., Thiel, V., Wiese, J., Labes, A., and Imhoff, J.F. (2013). Algae as an important environment for bacteria – phylogenetic relationships among new bacterial species isolated from algae. Phycologia *52*, 14–24.

Göker, M., Cleland, D., Saunders, E., Lapidus, A., Nolan, M., Lucas, S., Hammon, N., Deshpande, S., Cheng, J.-F., Tapia, R., et al. (2011). Complete genome sequence of Isosphaera pallida type strain (IS1B). Stand. Genomic Sci. *4*, 63–71.

Grant, J.R., Arantes, A.S., and Stothard, P. (2012). Comparing thousands of circular genomes using the CGView Comparison Tool. BMC Genomics *13*, 202.

Guo, M., Zhou, Q., Zhou, Y., Yang, L., Liu, T., Yang, J., Chen, Y., Su, L., Xu, J., Chen, J., et al. (2014). Genomic evolution of 11 type strains within family Planctomycetaceae. PLoS One *9*, e86752.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075.

Hempel, M., Blume, M., Blindow, I., and Gross, E.M. (2008). Epiphytic bacterial community composition on two common submerged macrophytes in brackish water and freshwater. BMC Microbiol. *8*, 58.

Hengst, M.B., Andrade, S., González, B., and Correa, J.A. (2010). Changes in epiphytic bacterial communities of intertidal seaweeds modulated by host, temporality, and copper enrichment. Microb. Ecol. *60*, 282–290.

Hieu, C.X., Voigt, B., Albrecht, D., Becher, D., Lombardot, T., Glöckner, F.O., Amann, R., Hecker, M., and Schweder, T. (2008). Detailed proteome analysis of growing cells of the planctomycete Rhodopirellula baltica SH1T. Proteomics *8*, 1608–1623.

Hou, S., Makarova, K.S., Saw, J.H.W., Senin, P., Ly, B. V, Zhou, Z., Ren, Y., Wang, J., Galperin, M.Y., Omelchenko, M. V, et al. (2008). Complete genome sequence of the extremely acidophilic methanotroph isolate V4, Methylacidiphilum infernorum, a representative of the bacterial phylum Verrucomicrobia. Biol. Direct *3*, 26.

Hu, Z., van Alen, T., Jetten, M.S.M., and Kartal, B. (2013). Lysozyme and penicillin inhibit the growth of anaerobic ammonium-oxidizing planctomycetes. Appl. Environ.

Microbiol. *79*, 7763–7769.

Huang, X., Ke, C., and Wang, W. (2010). Cadmium and copper accumulation and toxicity in the macroalga Gracilaria tenuistipitata. Aquat. Biol. *11*, 17–26.

Jenkins, C., and Fuerst, J.A. (2001). Phylogenetic analysis of evolutionary relationships of the planctomycete division of the domain bacteria based on amino acid sequences of elongation factor Tu. J. Mol. Evol. *52*, 405–418.

Jeske, O., Jogler, M., Petersen, J., Sikorski, J., and Jogler, C. (2013). From genome mining to phenotypic microarrays: Planctomycetes as source for novel bioactive molecules. Antonie Van Leeuwenhoek *104*, 551–567.

Jeske, O., Schüler, M., Schumann, P., Schneider, A., Boedeker, C., Jogler, M., Bollschweiler, D., Rohde, M., Mayer, C., Engelhardt, H., et al. (2015). Planctomycetes do possess a peptidoglycan cell wall. Nat. Commun. *6*, 7116.

Jetten, M. (1998). The anaerobic oxidation of ammonium. FEMS Microbiol. … *22*, 421–437.

Jetten, M. S. M., Camp, H.J.M., O. D., Kuenen, J. G., and S., and M. (2010). "Order II. 'Candidatus Brocadiales' ord. nov.," In Bac- Teroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobac- Teres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes, pp. 918–925.

Jogler, C., Glöckner, F.O., and Kolter, R. (2011). Characterization of Planctomyces limnophilus and development of genetic tools for its manipulation establish it as a model species for the phylum Planctomycetes. Appl. Environ. Microbiol. *77*, 5826–5829.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240.

Jun, S.-R., Sims, G.E., Wu, G.A., and Kim, S.-H. (2010). Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. Proc. Natl. Acad. Sci. U. S. A. *107*, 133–138.

Kalyuzhnyi, S. V., Shestakova, N.M., Tourova, T.P., Poltaraus, a. B., Gladchenko, M. a., Trukhina, a. I., and Nazina, T.N. (2010). Phylogenetic analysis of a microbial community involved in anaerobic oxidation of ammonium nitrogen. Microbiology *79*, 237–246.

Kartal, B., Kuenen, J.G., and van Loosdrecht, M.C.M. (2010). Engineering. Sewage treatment with anammox. Science *328*, 702–703.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics *28*, 1647–1649.

Klindworth, A., Richter, M., Richter-Heitmann, T., Wegner, C.-E., Frank, C.S., Harder, J., and Glöckner, F.O. (2014). Permanent draft genome of Rhodopirellula rubra SWK7. Mar. Genomics *13*, 11–12.

Kuffner, I., and Paul, V. (2001). Effects of nitrate, phosphate and iron on the growth of macroalgae and benthic cyanobacteria from Cocos Lagoon, Guam. Mar. Ecol. Prog. Ser. *222*, 63–72.

Kulichevskaya, I.S., Ivanova, A.O., Belova, S.E., Baulina, O.I., Bodelier, P.L.E., Rijpstra, W.I.C., Sinninghe Damsté, J.S., Zavarzin, G. a, and Dedysh, S.N. (2007). Schlesneria paludicola gen. nov., sp. nov., the first acidophilic member of the order Planctomycetales, from Sphagnum-dominated boreal wetlands. Int. J. Syst. Evol. Microbiol. *57*, 2680–2687.

Kulichevskaya, I.S., Ivanova, A.O., Baulina, O.I., Bodelier, P.L.E., Damsté, J.S.S., and Dedysh, S.N. (2008). Singulisphaera acidiphila gen. nov., sp. nov., a non-filamentous, Isosphaera-like planctomycete from acidic northern wetlands. Int. J. Syst. Evol. Microbiol. *58*, 1186–1193.

Kulichevskaya, I.S., Detkova, E.N., Bodelier, P.L.E., Rijpstra, W.I.C., Damsté, J.S.S., and Dedysh, S.N. (2012). Singulisphaera rosea sp. nov., a planctomycete from acidic Sphagnum peat, and emended description of the genus Singulisphaera. Int. J. Syst. Evol. Microbiol. *62*, 118–123.

Lachnit, T., Meske, D., Wahl, M., Harder, T., and Schmitz, R. (2011). Epibacterial community patterns on marine macroalgae are host-specific but temporally variable. Environ. Microbiol. *13*, 655–665.

Lage, O.M., and Bondoso, J. (2011). Planctomycetes diversity associated with macroalgae. Fems Micriobiology Ecol. *78*, 366–375.

Lage, O.M., and Bondoso, J. (2012). Bringing Planctomycetes into pure culture. Front. Microbiol. *3*, 405.

Lage, O.M., and Bondoso, J. (2014). Planctomycetes and macroalgae, a striking association. Front. Microbiol. *5*, 267.

Lage, O.M., Bondoso, J., and Lobo-da-Cunha, A. (2013). Insights into the ultrastructural morphology of novel Planctomycetes. Antonie Van Leeuwenhoek *104*, 467–476.

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M.R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., et al. (2015). Insights from 20 years of bacterial genome sequencing. Funct. Integr. Genomics *15*, 141–161.

Langó, Z. (2005). “Who has first observed planctomyces” (or data to the history of Planctomyces bekefii). Acta Microbiol. Immunol. Hung. *52*, 73–84.

Leaver, M., Domínguez-Cuevas, P., Coxhead, J.M., Daniel, R.A., and Errington, J. (2009). Life without a wall or division machine in Bacillus subtilis. Nature *457*, 849–853.

Lee, K.-C., Webb, R.I., Janssen, P.H., Sangwan, P., Romeo, T., Staley, J.T., and Fuerst, J.A. (2009). Phylum Verrucomicrobia representatives share a compartmentalized cell plan with members of bacterial phylum Planctomycetes. BMC Microbiol. *9*, 5.

Li, L., Stoeckert, C.J., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. *13*, 2178–2189.

Liechti, G.W., Kuru, E., Hall, E., Kalinda, A., Brun, Y. V, VanNieuwenhze, M., and Maurelli, A.T. (2014). A new metabolic cell-wall labelling method reveals peptidoglycan in Chlamydia trachomatis. Nature *506*, 507–510.

Liesack, W., König, H., Schlesner, H., and Hirsch, P. (1986). Chemical composition of the peptidoglycan-free cell envelopes of budding bacteria of the Pirella/Planctomyces group. Arch. Microbiol. *145*, 361–366.

Lindsay, M., Webb, R., Strous, M., Jetten,

M., Butler, M., Forde, R., and Fuerst, J. (2001). Cell compartmentalisation in planctomycetes: novel types of structural organisation for the bacterial cell. Arch. Microbiol. *175*, 413–429.

Liu, M., Dong, Y., Zhao, Y., Zhang, G., Zhang, W., and Xiao, T. (2010). Structures of bacterial communities on the surface of Ulva prolifera and in seawaters in an Ulva blooming region in Jiaozhou Bay, China. World J. Microbiol. Biotechnol. *27*, 1703–1712.

Liu, Y., Yao, T., Jiao, N., Kang, S., Zeng, Y., and Huang, S. (2006). Microbial community structure in moraine lakes and glacial meltwaters, Mount Everest. FEMS Microbiol. Lett. *265*, 98–105.

Longford, S., Tujula, N., Crocetti, G., Holmes, A., Holmström, C., Kjelleberg, S., Steinberg, P., and Taylor, M. (2007). Comparisons of diversity of bacterial communities associated with three sessile marine eukaryotes. Aquat. Microb. Ecol. *48*, 217–229.

Lonhienne, T.G.A., Sagulenko, E., Webb, R.I., Lee, K.-C., Franke, J., Devos, D.P., Nouwens, A., Carroll, B.J., and Fuerst, J.A. (2010). Endocytosis-like protein uptake in the bacterium Gemmata obscuriglobus. Proc. Natl. Acad. Sci. U. S. A. *107*, 12883–12888.

Lucheta, A.R., Otero, X.L., Macías, F., and Lambais, M.R. (2013). Bacterial and archaeal communities in the acid pit lake sediments of a chalcopyrite mine. Extremophiles *17*, 941–951.

Lugtenberg, B., and Van Alphen, L. (1983). Molecular architecture and functioning of the outer membrane of Escherichia coli and other gram-negative bacteria. Biochim. Biophys. Acta - Rev. Biomembr. *737*, 51–115.

Mäkinen, V., Salmela, L., and Ylinen, J. (2012). Normalized N50 assembly metric using gap-restricted co-linear chaining. BMC Bioinformatics *13*, 255.

Mavromatis, K., Land, M.L., Brettin, T.S., Quest, D.J., Copeland, A., Clum, A., Goodwin, L., Woyke, T., Lapidus, A., Klenk, H.P., et al. (2012). The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation. PLoS One *7*, e48837.

Mazaheri Nezhad Fard, R., Barton, M.D., and Heuzenroeder, M.W. (2011). Bacteriophage-mediated transduction of antibiotic resistance in enterococci. Lett. Appl. Microbiol. *52*, 559–564.

McCarren, J., and DeLong, E.F. (2007). Proteorhodopsin photosystem gene clusters exhibit co-evolutionary trends and shared ancestry among diverse marine microbial phyla. Environ. Microbiol. *9*, 846–858.

McInerney, J.O., Martin, W.F., Koonin, E. V, Allen, J.F., Galperin, M.Y., Lane, N., Archibald, J.M., and Embley, T.M. (2011). Planctomycetes and eukaryotes: a case of analogy not homology. Bioessays *33*, 810–817.

Meyer, F., Overbeek, R., and Rodriguez, A. (2009). FIGfams: yet another set of protein families. Nucleic Acids Res. *37*, 6643–6654.

Nasir, A., Naeem, A., Khan, M.J., Nicora, H.D.L., and Caetano-Anollés, G. (2011). Annotation of Protein Domains Reveals Remarkable Conservation in the Functional Make up of Proteomes Across Superkingdoms. Genes (Basel). *2*, 869–911.

Van Niftrik, L., Geerts, W.J.C., Van Donselaar, E.G., Humbel, B.M., Webb, R.I.,

Harhangi, H.R., Camp, H.J.M.O. Den, Fuerst, J.A., Verkleij, A.J., Jetten, M.S.M., et al. (2009). Cell division ring, a new cell division protein and vertical inheritance of a bacterial organelle in anammox planctomycetes. Mol. Microbiol. *73*, 1009–1019.

van Niftrik, L., van Helden, M., Kirchen, S., van Donselaar, E.G., Harhangi, H.R., Webb, R.I., Fuerst, J.A., Op den Camp, H.J.M., Jetten, M.S.M., and Strous, M. (2010). Intracellular localization of membrane-bound ATPases in the compartmentalized anammox bacterium "Candidatus Kuenenia stuttgartiensis". Mol. Microbiol. *77*, 701–715.

de Oliveira, L.S., Gregoracci, G.B., Silva, G.G.Z., Salgado, L.T., Filho, G.A., Alves-Ferreira, M., Pereira, R.C., and Thompson, F.L. (2012). Transcriptomic analysis of the red seaweed Laurencia dendroidea (Florideophyceae, Rhodophyta) and its microbiome. BMC Genomics *13*, 487.

Oren, A., and Papke, R.T. (2010). Molecular Phylogeny of Microorganisms (Horizon Scientific Press).

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J. V, Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. *33*, 5691–5702.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055.

Pearson, A., Budin, M., and Brocks, J.J. (2003). Phylogenetic and biochemical evidence for sterol synthesis in the bacterium Gemmata obscuriglobus. Proc. Natl. Acad. Sci. U. S. A. *100*, 15352–15357.

Peng, Y., Leung, H.C.M., Yiu, S.M., and Chin, F.Y.L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics *28*, 1420–1428.

Pilhofer, M., Rosati, G., Ludwig, W., Schleifer, K.-H., and Petroni, G. (2007). Coexistence of tubulins and ftsZ in different Prosthecobacter species. Mol. Biol. Evol. *24*, 1439–1442.

Pimentel-Elardo, S., Wehrl, M., Friedrich, A., Jensen, P., and Hentschel, U. (2003). Isolation of planctomycetes from Aplysina sponges. Aquat. Microb. Ecol. *33*, 239–245.

Pollet, T., Humbert, J.-F., and Tadonléké, R.D. (2014). Planctomycetes in lakes: poor or strong competitors for phosphorus? Appl. Environ. Microbiol. *80*, 819–828.

Rath, D., Amlinger, L., Rath, A., and Lundgren, M. (2015). The CRISPR-Cas immune system: Biology, mechanisms and applications. Biochimie *117*, 119–128.

Reynaud, E.G., and Devos, D.P. (2011). Transitional forms between the three domains of life and evolutionary implications. Proc. Biol. Sci. *278*, 3321–3328.

Richter, M., Richter-Heitmann, T., Klindworth, A., Wegner, C.-E., Frank, C.S., Harder, J., and Glöckner, F.O. (2014). Permanent draft genomes of the Rhodopirellula maiorica strain SM1. Mar. Genomics *13*, 19–20.

Richter-Heitmann, T., Richter, M., Klindworth, A., Wegner, C.-E., Frank, C.S., Glöckner, F.O., and Harder, J. (2014). Permanent draft genomes of the two Rhodopirellula europaea strains 6C and SH398. Mar. Genomics *13*, 15–16.

Santarella-Mellwig, R., Franke, J., Jaedicke, A., Gorjanacz, M., Bauer, U., Budd, A., Mattaj, I.W., and Devos, D.P. (2010a). The compartmentalized bacteria of the planctomycetes-verrucomicrobia-chlamydiae superphylum have membrane coat-like proteins. PLoS Biol. *8*.

Santarella-Mellwig, R., Franke, J., Jaedicke, A., Gorjanacz, M., Bauer, U., Budd, A., Mattaj, I.W., and Devos, D.P. (2010b). The compartmentalized bacteria of the planctomycetes-verrucomicrobia-chlamydiae superphylum have membrane coat-like proteins. PLoS Biol. *8*, e1000281.

Santarella-Mellwig, R., Pruggnaller, S., Roos, N., Mattaj, I.W., and Devos, D.P. (2013). Three-Dimensional Reconstruction of Bacteria with a Complex Endomembrane System. PLoS Biol. *11*, e1001565.

Schlesner, H. (1989). Planctomyces brasiliensis sp. nov., a Halotolerant Bacterium from a Salt Pit. Syst. Appl. Microbiol. *12*, 159–161.

Schlesner, H., and Hirsch, P. (1984). Assignment of ATCC 27377 to Pirella gen. nov. as Pirella staleyi comb. nov. Int. J. Syst. Bacteriol. *34*, 492–495.

Schlesner, H., and Hirsch, P. (1987). Rejection of the Genus Name Pirella for Pear-Shaped Budding Bacteria and Proposal to Create the Genus Pirellula gen. nov. Int. J. Syst. Bacteriol. *37*, 441–441.

Schmid, M., Walsh, K., Webb, R., Rijpstra, W.I.C., van de Pas-Schoonen, K., Verbruggen, M.J., Hill, T., Moffett, B., Fuerst, J., Schouten, S., et al. (2003). Candidatus “Scalindua brodae”, sp. nov., Candidatus “Scalindua wagneri”, sp. nov., two new species of anaerobic ammonium oxidizing bacteria. Syst. Appl. Microbiol. *26*, 529–538.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069.

Sheldon, R.A. (2011). Characteristic features and biotechnological applications of cross-linked enzyme aggregates (CLEAs). Appl. Microbiol. Biotechnol. *92*, 467–477.

Speth, D.R., van Teeseling, M.C.F., and Jetten, M.S.M. (2012). Genomic analysis indicates the presence of an asymmetric bilayer outer membrane in planctomycetes and verrucomicrobia. Front. Microbiol. *3*, 304.

Staley, J.T. (1973). Budding bacteria of the Pasteuria-Blastobacter group. Can. J. Microbiol. *19*, 609–614.

Staley, J.T., and Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu. Rev. Microbiol. *39*, 321–346.

Starr, M.P., Sayre, R.M., and Schmidt, J.M. (1983). Assignment of ATCC 27377 to Planctomyces staleyi sp. nov. and Conservation of Pasteuria ramosa Metchnikoff 1888 on the Basis of Type Descriptive Material: Request for an Opinion. Int. J. Syst. Bacteriol. *33*, 666–671.

Strous, M., Fuerst, J.A., Kramer, E.H., Logemann, S., Muyzer, G., van de Pas-Schoonen, K.T., Webb, R., Kuenen, J.G., and Jetten, M.S. (1999). Missing lithotroph identified as new planctomycete. Nature *400*, 446–449.

Strous, M., Rijpstra, W.I.C., and Damste, J.S.S. (2002). Linearly concatenated cyclobutane lipids form a dense bacterial membrane. *419*, 8–12.

Strous, M., Kraft, B., Bisdorf, R., and Tegetmeyer, H.E. (2012). The binning of

metagenomic contigs for microbial physiology of mixed cultures. Front. Microbiol. *3*, 410.

Sutcliffe, I.C. (2010). A phylum level perspective on bacterial cell envelope architecture. Trends Microbiol. *18*, 464–470.

Teeling, H., Lombardot, T., Bauer, M., Ludwig, W., and Glöckner, F.O. (2004). Evaluation of the phylogenetic position of the planctomycete "Rhodopirellula baltica" SH 1 by means of concatenated ribosomal protein sequences, DNA-directed RNA polymerase subunit sequences and whole genome trees. Int. J. Syst. Evol. Microbiol. *54*, 791–801.

van Teeseling, M.C.F., Mesman, R.J., Kuru, E., Espaillat, A., Cava, F., Brun, Y. V, VanNieuwenhze, M.S., Kartal, B., and van Niftrik, L. (2015). Anammox Planctomycetes have a peptidoglycan cell wall. Nat. Commun. *6*, 6878.

Tekere, M., Lötter, A., Olivier, J., Jonker, N., and Venter, S. (2013). Metagenomic analysis of bacterial diversity of Siloam hot water spring, Limpopo, South Africa. African J. Biotechnol. *10*, 18005–18012.

Toledo-Ortiz, G., Huq, E., and Rodríguez-Concepción, M. (2010). Direct regulation of phytoene synthase gene expression and carotenoid biosynthesis by phytochrome-interacting factors. Proc. Natl. Acad. Sci. U. S. A. *107*, 11626–11631.

Tujula, N.A., Crocetti, G.R., Burke, C., Thomas, T., Holmström, C., and Kjelleberg, S. (2010). Variability and abundance of the epiphytic bacterial community associated with a green marine Ulvacean alga. ISME J. *4*, 301–311.

Wagner, M., and Horn, M. (2006). The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. Curr. Opin. Biotechnol. *17*, 241–249.

Wagner-Döbler, I., Beil, W., Lang, S., Meiners, M., and Laatsch, H. (2002). Integrated approach to explore the potential of marine microorganisms for the production of bioactive metabolites. Adv. Biochem. Eng. Biotechnol. *74*, 207–238.

Ward, N.L. (2010). "Family I. Planctomycetaceae Schlesner and Stackebrandt 1987, 179VP (Effec- tive publication: Schlesner and Stackebrandt 1986, 175) emend. Ward (this volume),." In In The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes, 2nd Edn, pp. 879–925.

Ward, N., Staley, J., and Fuerst, J. (2006). The order Planctomycetales, including the genera Planctomyces, Pirellula, Gemmata and Isosphaera and the Candidatus genera Brocadia, Kuenenia and. In The Prokaryotes, pp. 757–793.

Ward, N.L., Rainey, F.A., Hedlund, B.P., Staley, J.T., Ludwig, W., and Stackebrandt, E. (2000). Comparative phylogenetic analyses of members of the order Planctomycetales and the division Verrucomicrobia: 23S rRNA gene sequence analysis supports the 16S rRNA gene sequence-derived phylogeny. Int. J. Syst. Evol. Microbiol. *50*, 1965–1972.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737–738.

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H.U., Bruccoleri, R., Lee, S.Y., Fischbach, M.A., Muller, R., Wohlleben, W., et al. (2015). antiSMASH 3.0--a

comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. gkv437 – .

Webster, N.S., and Bourne, D. (2007). Bacterial community structure associated with the Antarctic soft coral, Alcyonium antarcticum. FEMS Microbiol. Ecol. *59*, 81–94.

Webster, N.S., Wilson, K.J., Blackall, L.L., and Hill, R.T. (2001). Phylogenetic diversity of bacteria associated with the marine sponge Rhopaloeides odorabile. Appl. Environ. Microbiol. *67*, 434–444.

Wegner, C.-E., Richter-Heitmann, T., Klindworth, A., Klockow, C., Richter, M., Achstetter, T., Glöckner, F.O., and Harder, J. (2013). Expression of sulfatases in Rhodopirellula baltica and the diversity of sulfatases in the genus Rhodopirellula. Mar. Genomics *9*, 51–61.

Wegner, C.-E., Richter, M., Richter-Heitmann, T., Klindworth, A., Frank, C.S., Glöckner, F.O., and Harder, J. (2014). Permanent draft genome of Rhodopirellula sallentina SM41. Mar. Genomics *13*, 17–18.

Willey, J.M., and van der Donk, W.A. (2007). Lantibiotics: peptides of diverse structure and function. Annu. Rev. Microbiol. *61*, 477–501.

Winkelmann, N., and Harder, J. (2009). An improved isolation method for attached-living Planctomycetes of the genus Rhodopirellula. J. Microbiol. Methods *77*, 276–284.

Winkelmann, N., Jaekel, U., Meyer, C., Serrano, W., Rachel, R., Rosselló-Mora, R., and Harder, J. (2010). Determination of the diversity of Rhodopirellula isolates from European seas by multilocus sequence analysis. Appl. Environ. Microbiol. *76*, 776–785.

Woese, C.R. (1987). Bacterial evolution. Microbiol. Rev. *51*, 221–271.

Yamada, Y., Kuzuyama, T., Komatsu, M., Shin-Ya, K., Omura, S., Cane, D.E., and Ikeda, H. (2015). Terpene synthases are widely distributed in bacteria. Proc. Natl. Acad. Sci. U. S. A. *112*, 857–862.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

Zhang, W., Wu, X., Liu, G., Chen, T., Zhang, G., Dong, Z., Yang, X., and Hu, P. (2013). Pyrosequencing Reveals Bacterial Diversity in the Rhizosphere of Three Phragmites australis Ecotypes. Geomicrobiol. J. *30*, 593–599.

Zhang, Y., Du, B.-H., Jin, Z., Li, Z., Song, H., and Ding, Y.-Q. (2010). Analysis of bacterial communities in rhizosphere soil of healthy and diseased cotton (Gossypium sp.) at different plant growth stages. Plant Soil *339*, 447–455.

Zhou, Y., Liang, Y., Lynch, K.H., Dennis, J.J., and Wishart, D.S. (2011). PHAST: a fast phage search tool. Nucleic Acids Res. *39*, W347–W352.

Appendix

Appendix I. Groups of clustered proteins between LF1 + UC8 + FC18 retrived with OrthoMCL, InterProScan and PSI-BLAST.

Cluster	NAME / ID	Suggested protein	Protein Length (aa)	TMHs (TMHMM)	Sig. Peptide
	LF1_03573 hypothetical protein	Glycylpeptide N-tetradecanoyltransferase	232	1,0	1-24 (0.640)
	UC8_02076 hypothetical protein	Methionine--tRNA ligase (EC 6.1.1.10) (Methionyl-tRNA synthetase) (MetRS)	667	0,0	NO
Prot 287	FC18_00481 Endoglucanase Z precursor		755	0,0	1-21 (0.629)
	FC18_01185 hypothetical protein	DnaJ-like protein MG200 homolog	300	0,0	NO
	FC18_04987 hypothetical protein	Mediator of RNA polymerase II transcription subunit 13 (Mediator complex subunit 13)	787	0,0	NO
	FC18_04018 Fungalysin metallopeptidase (M36)		1198	0,0	NO
	LF1_03994 hypothetical protein	Hexagonally packed intermediate-layer surface protein	460	0,0	NO
	UC8_01625 hypothetical protein	Chaperone protein ClpB	509	0,0	NO
	UC8_02056 hypothetical protein	Protease HtpX homolog (EC 3.4.24.-)	505	0,0	1-30 (0.529)
Prot 1497					
	FC18_06768 hypothetical protein	tRNA (guanine-N(7)-)-methyltransferase (EC 2.1.1.33) (tRNA (guanine(46)-N(7))-methyltransferase) (tRNA(m7G46)-methyltransferase)	511	0,0	1-38 (0.677)
	FC18_04625 hypothetical protein	Protease HtpX homolog (EC 3.4.24.-)	504	0,0	NO
	LF1_03996 ECF RNA polymerase sigma-E factor		172	0,0	NO
	UC8_01627 RNA polymerase sigma factor		219	0,0	NO
Prot1498	UC8_05320 ECF RNA polymerase sigma factor SigE		160	0,0	NO
	FC18_02770 RNA polymerase sigma factor RpoE		176	0,0	NO
	FC18_04627 RNA polymerase sigma factor SigV		168	0,0	NO
	LF1_00937 Cation efflux system protein CusA		1189	13,0	NO
Prot1537	UC8_01695 Cation efflux system protein CusA		1325	14,0	NO
	UC8_02386 Cation efflux system protein		1166	12,0	NO

Cluster	NAME / ID	Suggested protein	Protein Length (aa)	TMHs (TMHMM)	Sig. Peptide
Prot 2272	UC8_01736 Cation efflux system protein CusA		1171	13,0	NO
	FC18_01420 Cation efflux system protein CusA		1212	14,0	NO
	LF1_00927 putative cadmium-transporting ATPase		854	6,0	NO
	UC8_01731 putative cadmium-transporting ATPase		821	6,0	NO
	UC8_02216 putative cadmium-transporting ATPase		813	5,0	NO
	FC18_00727 putative cadmium-transporting ATPase		833	6,0	NO
Prot 2273	LF1_00940 Outer membrane efflux protein		584	0,0	1-38 (0.586)
	UC8_01734 Outer membrane efflux protein		543	0,0	1-38 (0.591)
	UC8_02382 Outer membrane efflux protein		529	0,0	1-42 (0.496)
	FC18_01412 Outer membrane efflux protein		497	0,0	NO
Prot 2274	LF1_00938 Cation efflux system protein CusB precursor		707	1,0	NO
	UC8_01735 Cation efflux system protein CusB precursor		707	1,0	NO
	UC8_02384 Cation efflux system protein CusB precursor		741	1,0	NO
	FC18_01419 Cation efflux system protein CusB precursor		758	1,0	NO
Prot 2762	LF1_00936 Archaeal TRASH domain protein		384	0,0	1-26 (0.724)
	UC8_01737 Archaeal TRASH domain protein		459	0,0	1-26 (0.811)
	UC8_02387 Archaeal TRASH domain protein		396	0,0	1-21 (0.767)
	FC18_01421 Archaeal TRASH domain protein		275	0,0	1-19 (0.593)
Prot 2870	LF1_04878 hypothetical protein	Protease HtpX (EC 3.4.24.-) (Heat shock protein HtpX)	666	4,0	NO
	LF1_03767 heat shock protein HtpX		690	3,0	NO
	UC8_00524 hypothetical protein	Uncharacterised protein VP1481 - via uniprot	613	3,0	NO
	FC18_04575 hypothetical protein	Protease HtpX homolog (EC 3.4.24.-)	326	4,0	NO
Prot 3164	LF1_05265 Methyl-accepting chemotaxis protein McpC		280	0,0	NO
	UC8_01791 Methyl-accepting chemotaxis protein 4		467	2,0	1-37 (0.467)
	FC18_03302 Methyl-accepting chemotaxis protein McpC		457	2,0	1-37 (0.548)
Prot 3500	LF1_01379 ADP-L-glycero-D-manno-heptose-6-epimerase		334	0,0	NO
	UC8_04981 NAD dependent epimerase/dehydratase family protein		387	0,0	NO
	FC18_00088 NAD dependent epimerase/dehydratase family protein		381	0,0	NO
Prot 3501	LF1_01377 Sulfotransferase domain protein		298	0,0	NO
	UC8_03890 Sulfotransferase domain protein		295	0,0	NO

Cluster	NAME / ID	Suggested protein	Protein Length (aa)	TMHs (TMHMM)	Sig. Peptide
	FC18_00098 Sulfotransferase domain protein		320	0,0	NO
	LF1_03200 Fructosamine kinase		309	0,0	NO
Prot 3522		Ferric uptake regulation protein 2 (Ferric uptake regulator 2)			
	UC8_00862 hypothetical protein		307	0,0	NO
	FC18_00736 Fructosamine kinase		273	0,0	NO
	LF1_01670 NAD(P)-specific glutamate dehydrogenase		451	0,0	NO
Prot 3523	UC8_03720 NADP-specific glutamate dehydrogenase		449	0,0	NO
	FC18_00754 NADP-specific glutamate dehydrogenase		442	0,0	NO
	LF1_04485 Endonuclease/Exonuclease/phosphatase family protein		536	0,0	1-27 (0.574)
Prot 3527	UC8_02802 Endonuclease/Exonuclease/phosphatase family protein		605	0,0	NO
	FC18_00867 Endonuclease/Exonuclease/phosphatase family protein		539	1,0	1-39 (0.586)
	LF1_04491 hypothetical protein	UPF0507 protein SCY_4172	458	0,0	NO
Prot 3545	UC8_00364 hypothetical protein	Long-chain-fatty-acid--CoA ligase bubblegum-like (EC 6.2.1.3)	458	0,0	NO
	FC18_01292 hypothetical protein	Putative DEAD-box ATP-dependent RNA helicase 29 (EC 3.6.4.13)	343	0,0	NO
	LF1_00925 hypothetical protein	Gag polyprotein (Pr55Gag) [Cleaved into: Matrix protein p17 (MA); Capsid protein p24 (CA); Spacer peptide 1 (SP1) (p2); Nucleocapsid protein p7 (NC); Spacer peptide 2 (SP2) (p1); p6-gag]	114	1,0	1-25 (0.778)
Prot 3547					
	UC8_01725 hypothetical protein	Gag polyprotein (Pr55Gag) [Cleaved into: Matrix protein p17 (MA); Capsid protein p24 (CA); Spacer peptide 1 (SP1) (p2); Nucleocapsid protein p7 (NC); Spacer peptide 2 (SP2) (p1); p6-gag]	222	0,0	1-22 (0.703)
	FC18_01394 hypothetical protein	Putative membrane protein YdgH	343	0,0	NO
	LF1_04721 Low molecular weight protein-tyrosine-phosphatase YfkJ		165	0,0	NO
Prot 3550	UC8_02929 Low molecular weight protein-		154	0,0	NO

Cluster	NAME / ID	Suggested protein	Protein Length (aa)	TMHs (TMHMM)	Sig. Peptide
	tyrosine-phosphatase YfkJ				
	FC18_01448 Low molecular weight protein-tyrosine-phosphatase YfkJ		165	0,0	NO
	LF1_02338 hypothetical protein	UPF0160 protein	238	2,0	NO
	UC8_01429 hypothetical protein	UPF0160 protein C27H6.8	236	2,0	NO
Prot 3563		UDP-N-acetylenolpyruvoylglucosamine reductase (EC 1.3.1.98) (UDP-N-acetylmuramate dehydrogenase)			
	FC18_01761 hypothetical protein		239	2,0	NO
	FC18_02515 hypothetical protein		249	0	NO
Prot 3581	LF1_01941 hypothetical protein	Uncharacterised protein YqjF	248	0	NO
	UC8_02455 hypothetical protein		227	0	NO
	LF1_03590 Chaperone protein DnaJ		186	0,0	NO
Prot 3592	UC8_02572 Curved DNA-binding protein		195	0,0	NO
	FC18_02713 Chaperone protein DnaJ		189	0,0	NO
	LF1_04087 O-Antigen ligase		883	11,0	NO
Prot 3598	UC8_04306 O-Antigen ligase		884	11,0	NO
	FC18_02868 O-Antigen ligase		851	12,0	NO
	LF1_00530 Prolyl tripeptidyl peptidase precursor		795	0,0	1-45 (0.500)
Prot 3599	UC8_01779 Prolyl tripeptidyl peptidase precursor		757	0,0	NO
	FC18_02873 Prolyl tripeptidyl peptidase precursor		1228	0,0	1-23 (0.819)
	LF1_00086 transaldolase/EF-hand domain-containing protein		240	0,0	1-25 (0.826)
Prot 3601	UC8_05663 EF hand		225	0,0	1-25 (0.663)
	FC18_02998 EF hand		215	0,0	1-22 (0.589)
	LF1_01721 BlaR1 peptidase M56		402	3,0	NO
Prot 3621	UC8_04532 Regulatory protein BlaR1		636	4,0	NO
	FC18_03421 Regulatory protein BlaR1		973	4,0	NO
	LF1_02193 hypothetical protein	Protein ea22	291	0,0	NO
Prot 3629	UC8_04049 hypothetical protein	Protein ea22	291	0,0	NO
	FC18_03613 hypothetical protein	UPF0160 protein	292	0,0	NO
	LF1_02063 Amino acid permease		731	11,0	NO
Prot 3636	UC8_02927 Amino acid permease		733	11,0	NO
	FC18_03777 Amino acid permease		755	12,0	NO
	LF1_04991 Capsid protein (F protein)		97	0,0	NO
Prot 3646	UC8_05696 Capsid protein (F protein)		427	0,0	NO
	FC18_03945 Capsid protein (F protein)		427	0,0	NO
	LF1_04986 Bacteriophage scaffolding protein D		152	0,0	NO
Prot 3647	UC8_05697 Bacteriophage scaffolding protein D		152	0,0	NO
	FC18_03947 Bacteriophage scaffolding protein D		152	0,0	NO
	LF1_00130 Glycosyl transferases group 1		373	0,0	NO
Prot 3653	UC8_00838 Glycosyl transferases group 1		356	0,0	NO
	FC18_04128 Glycosyl transferases group 1		385	0,0	NO
	LF1_05256 hypothetical protein	Transmembrane protein 143	182	0,0	NO
Prot 3655	UC8_01344 hypothetical protein	Transmembrane protein 143	182	0,0	NO
	FC18_04244 hypothetical protein	Protein ea22	182	0,0	NO

Cluster	NAME / ID	Suggested protein	Protein Length (aa)	TMHs (TMHMM)	Sig. Peptide
Prot 3661	LF1_03699 NTE family protein RssA		316	0,0	NO
	UC8_05461 NTE family protein RssA		313	0,0	NO
	FC18_04405 NTE family protein RssA		607	0,0	NO
Prot 3662	LF1_04665 Mannosylfructose-phosphate phosphatase		269	0,0	NO
	UC8_02514 Mannosylfructose-phosphate phosphatase		276	0,0	NO
	FC18_04428 Kanosamine-6-phosphate phosphatase		277	0,0	NO
Prot 3668	LF1_03117 hypothetical protein	Electron transport complex subunit D	419	1,0	NO
	UC8_04006 hypothetical protein	Sickle tail protein (Enhancer trap locus 4)	452	1,0	NO
	FC18_04553 hypothetical protein	Transmembrane protein 143	482	2,0	NO
Prot 3676	LF1_05134 Beta-lactamase TEM precursor		286	0,0	1-25 (0.483)
	UC8_01336 Beta-lactamase TEM precursor		286	0,0	1-25 (0.483)
	FC18_04720 Beta-lactamase TEM precursor		160	0,0	NO
Prot 3677	LF1_01294 hypothetical protein	Probable iron export permease protein FetB	195	4,0	NO
	UC8_00732 hypothetical protein	Membrane protein insertase YidC (Foldase YidC) (Membrane integrase YidC) (Membrane protein YidC)	204	3,0	NO
	FC18_04738 hypothetical protein	Envelope glycoprotein E (gE)	190	3,0	NO
Prot 3678	LF1_00563 putative ABC transporter ATP-binding protein YbbL		226	0,0	NO
	UC8_00146 putative ABC transporter ATP-binding protein YbbL		254	0,0	NO
	FC18_04744 L-cystine import ATP-binding protein TcyN		217	0,0	NO
Prot 3679	LF1_00564 hypothetical protein	3-phosphoshikimate 1-carboxyvinyltransferase (EC 2.5.1.19) (5-enolpyruvylshikimate-3-phosphate synthase) (EPSP synthase) (EPSPS)	257	6,0	NO
	FC18_04745 hypothetical protein	Probable iron export permease protein FetB	264	7,0	NO

Appendix II. GO terms of the shared proteins among LF1, UC8 and FC18 retrieved with Blast2GO.

Level	GO ID	Term	Type	#Seqs
1	GO:0003674	molecular_function	Molecular Function	72
2	GO:0060089	molecular transducer activity	Molecular Function	3
2	GO:0005198	structural molecule activity	Molecular Function	3
2	GO:0001071	nucleic acid binding transcription factor activity	Molecular Function	5
2	GO:0000988	transcription factor activity, protein binding	Molecular Function	5
2	GO:0005488	binding	Molecular Function	29
2	GO:0005215	transporter activity	Molecular Function	16
2	GO:0003824	catalytic activity	Molecular Function	37
3	GO:0036094	small molecule binding	Molecular Function	7
3	GO:0022892	substrate-specific transporter activity	Molecular Function	4
3	GO:0016491	oxidoreductase activity	Molecular Function	3
3	GO:0097367	carbohydrate derivative binding	Molecular Function	3
3	GO:0004871	signal transducer activity	Molecular Function	3
3	GO:0000990	transcription factor activity, core RNA polymerase binding	Molecular Function	5
3	GO:0005515	protein binding	Molecular Function	4
3	GO:1901363	heterocyclic compound binding	Molecular Function	12
3	GO:0043167	ion binding	Molecular Function	14
3	GO:0003700	transcription factor activity, sequence-specific DNA binding	Molecular Function	5
3	GO:0022857	transmembrane transporter activity	Molecular Function	4
3	GO:0097159	organic cyclic compound binding	Molecular Function	12
3	GO:0048037	cofactor binding	Molecular Function	3
3	GO:0016740	transferase activity	Molecular Function	2
3	GO:0016787	hydrolase activity	Molecular Function	26
4	GO:0022804	active transmembrane transporter activity	Molecular Function	4
4	GO:1901265	nucleoside phosphate binding	Molecular Function	7
4	GO:0022891	substrate-specific transmembrane transporter activity	Molecular Function	4
4	GO:0016638	oxidoreductase activity, acting on the CH-NH2 group of donors	Molecular Function	3
4	GO:0008233	peptidase activity	Molecular Function	8
4	GO:0001882	nucleoside binding	Molecular Function	3
4	GO:0017171	serine hydrolase activity	Molecular Function	4
4	GO:0016788	hydrolase activity, acting on ester bonds	Molecular Function	6
4	GO:0043169	cation binding	Molecular Function	11
4	GO:0000996	core DNA-dependent RNA polymerase binding	Molecular Function	5

Level	GO ID	Term	Type	#Seqs
		promoter specificity activity		
4	GO:0043168	anion binding	Molecular Function	3
4	GO:0016798	hydrolase activity, acting on glycosyl bonds	Molecular Function	2
4	GO:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	Molecular Function	3
4	GO:0016817	hydrolase activity, acting on acid anhydrides	Molecular Function	7
4	GO:0003676	nucleic acid binding	Molecular Function	5
4	GO:0050662	coenzyme binding	Molecular Function	3
4	GO:0016782	transferase activity, transferring sulfur-containing groups	Molecular Function	2
5	GO:0015399	primary active transmembrane transporter activity	Molecular Function	4
5	GO:0015075	ion transmembrane transporter activity	Molecular Function	4
5	GO:0008146	sulfotransferase activity	Molecular Function	2
5	GO:0042578	phosphoric ester hydrolase activity	Molecular Function	3
5	GO:0035639	purine ribonucleoside triphosphate binding	Molecular Function	3
5	GO:0016639	oxidoreductase activity, acting on the CH-NH2 group of donors, NAD or NADP as acceptor	Molecular Function	3
5	GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds	Molecular Function	2
5	GO:0046872	metal ion binding	Molecular Function	11
5	GO:0016987	sigma factor activity	Molecular Function	5
5	GO:0000166	nucleotide binding	Molecular Function	7
5	GO:0001883	purine nucleoside binding	Molecular Function	3
5	GO:0032549	ribonucleoside binding	Molecular Function	3
5	GO:0004518	nuclease activity	Molecular Function	3
5	GO:0016812	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in cyclic amides	Molecular Function	3
5	GO:0016818	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	Molecular Function	7
5	GO:0003677	DNA binding	Molecular Function	5
5	GO:0070011	peptidase activity, acting on L-amino acid peptides	Molecular Function	8
5	GO:0016820	hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances	Molecular Function	4
6	GO:0004536	deoxyribonuclease activity	Molecular Function	3
6	GO:0005509	calcium ion binding	Molecular Function	4
6	GO:0008236	serine-type peptidase activity	Molecular Function	4
6	GO:0008238	exopeptidase activity	Molecular Function	1
6	GO:0008237	metallopeptidase activity	Molecular Function	6
6	GO:0032553	ribonucleotide binding	Molecular Function	3
6	GO:0046914	transition metal ion binding	Molecular Function	4

Level	GO ID	Term	Type	#Seqs
6	GO:0032550	purine ribonucleoside binding	Molecular Function	3
6	GO:0008324	cation transmembrane transporter activity	Molecular Function	4
6	GO:0008800	beta-lactamase activity	Molecular Function	3
6	GO:0016791	phosphatase activity	Molecular Function	3
6	GO:0015405	P-P-bond-hydrolysis-driven transmembrane transporter activity	Molecular Function	4
6	GO:0004175	endopeptidase activity	Molecular Function	7
6	GO:0017076	purine nucleotide binding	Molecular Function	3
6	GO:0016462	pyrophosphatase activity	Molecular Function	7
7	GO:0004180	carboxypeptidase activity	Molecular Function	1
7	GO:0004222	metalloendopeptidase activity	Molecular Function	5
7	GO:0008270	zinc ion binding	Molecular Function	4
7	GO:0008235	metalloexopeptidase activity	Molecular Function	1
7	GO:0032555	purine ribonucleotide binding	Molecular Function	3
7	GO:0004252	serine-type endopeptidase activity	Molecular Function	3
7	GO:0004721	phosphoprotein phosphatase activity	Molecular Function	3
7	GO:0017111	nucleoside-triphosphatase activity	Molecular Function	7
7	GO:0030554	adenyl nucleotide binding	Molecular Function	3
8	GO:0016887	ATPase activity	Molecular Function	7
8	GO:0004181	metallocarboxypeptidase activity	Molecular Function	1
8	GO:0032559	adenyl ribonucleotide binding	Molecular Function	3
8	GO:0004725	protein tyrosine phosphatase activity	Molecular Function	3
9	GO:0042623	ATPase activity, coupled	Molecular Function	4
9	GO:0005524	ATP binding	Molecular Function	3
10	GO:0043492	ATPase activity, coupled to movement of substances	Molecular Function	4
11	GO:0042626	ATPase activity, coupled to transmembrane movement of substances	Molecular Function	4
12	GO:0042625	ATPase activity, coupled to transmembrane movement of ions	Molecular Function	4
13	GO:0019829	cation-transporting ATPase activity	Molecular Function	4
1	GO:0005575	cellular_component	Cellular Component	20
2	GO:0019012	virion	Cellular Component	3
2	GO:0016020	membrane	Cellular Component	17
2	GO:0005576	extracellular region	Cellular Component	1
3	GO:0044421	extracellular region part	Cellular Component	1
3	GO:0044423	virion part	Cellular Component	3
3	GO:0044425	membrane part	Cellular Component	4
4	GO:0005615	extracellular space	Cellular Component	1
4	GO:0019028	viral capsid	Cellular Component	3

Level	GO ID	Term	Type	#Seqs
4	GO:0031224	intrinsic component of membrane	Cellular Component	4
5	GO:0016021	integral component of membrane	Cellular Component	4
1	GO:0008150	biological_process	Biological Process	56
2	GO:0065007	biological regulation	Biological Process	8
2	GO:0023052	signaling	Biological Process	3
2	GO:0044699	single-organism process	Biological Process	20
2	GO:0051704	multi-organism process	Biological Process	3
2	GO:0022610	biological adhesion	Biological Process	1
2	GO:0008152	metabolic process	Biological Process	33
2	GO:0051179	localization	Biological Process	17
2	GO:0071840	cellular component organization or biogenesis	Biological Process	3
2	GO:0050896	response to stimulus	Biological Process	6
2	GO:0009987	cellular process	Biological Process	27
3	GO:0044710	single-organism metabolic process	Biological Process	9
3	GO:1902578	single-organism localization	Biological Process	8
3	GO:0051716	cellular response to stimulus	Biological Process	3
3	GO:0071704	organic substance metabolic process	Biological Process	30
3	GO:0044238	primary metabolic process	Biological Process	27
3	GO:0044237	cellular metabolic process	Biological Process	17
3	GO:0044085	cellular component biogenesis	Biological Process	3
3	GO:0042221	response to chemical	Biological Process	3
3	GO:0044764	multi-organism cellular process	Biological Process	3
3	GO:0044763	single-organism cellular process	Biological Process	13
3	GO:0006807	nitrogen compound metabolic process	Biological Process	14
3	GO:0016043	cellular component organization	Biological Process	3
3	GO:0044419	interspecies interaction between organisms	Biological Process	3
3	GO:0009056	catabolic process	Biological Process	9
3	GO:0007155	cell adhesion	Biological Process	1
3	GO:0009058	biosynthetic process	Biological Process	5
3	GO:0044700	single organism signaling	Biological Process	3
3	GO:0051234	establishment of localization	Biological Process	17
3	GO:0050789	regulation of biological process	Biological Process	8
4	GO:0044712	single-organism catabolic process	Biological Process	3
4	GO:0044281	small molecule metabolic process	Biological Process	6
4	GO:0050794	regulation of cellular process	Biological Process	8
4	GO:0006725	cellular aromatic compound metabolic process	Biological Process	8
4	GO:1901360	organic cyclic compound metabolic process	Biological Process	11
4	GO:0046483	heterocycle metabolic process	Biological Process	11
4	GO:0019222	regulation of metabolic process	Biological Process	5

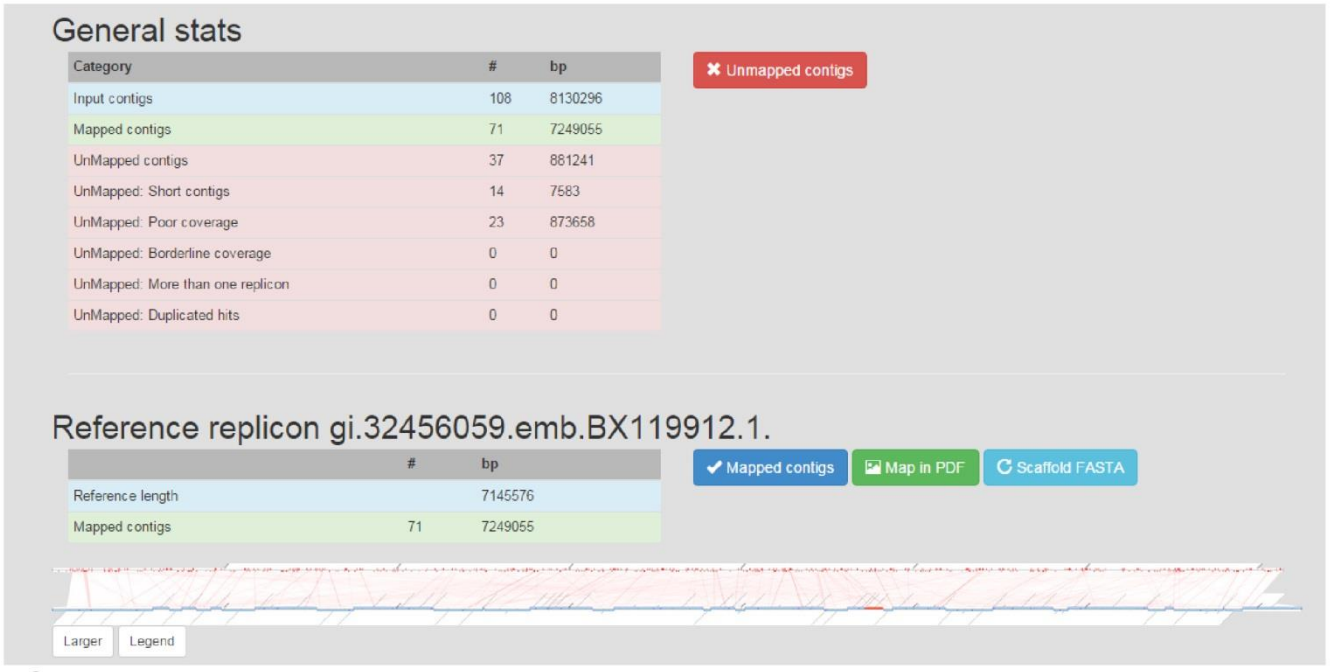
Level	GO ID	Term	Type	#Seqs
4	GO:0006793	phosphorus metabolic process	Biological Process	3
4	GO:0005975	carbohydrate metabolic process	Biological Process	3
4	GO:0055114	oxidation-reduction process	Biological Process	3
4	GO:0098609	cell-cell adhesion	Biological Process	1
4	GO:0022607	cellular component assembly	Biological Process	3
4	GO:0009636	response to toxic substance	Biological Process	3
4	GO:0017144	drug metabolic process	Biological Process	3
4	GO:0044249	cellular biosynthetic process	Biological Process	5
4	GO:0044403	symbiosis, encompassing mutualism through parasitism	Biological Process	3
4	GO:0006810	transport	Biological Process	17
4	GO:0044248	cellular catabolic process	Biological Process	6
4	GO:0007154	cell communication	Biological Process	3
4	GO:1901564	organonitrogen compound metabolic process	Biological Process	6
4	GO:0043170	macromolecule metabolic process	Biological Process	21
4	GO:0034641	cellular nitrogen compound metabolic process	Biological Process	11
4	GO:1901576	organic substance biosynthetic process	Biological Process	5
4	GO:1901575	organic substance catabolic process	Biological Process	9
4	GO:0006629	lipid metabolic process	Biological Process	3
5	GO:0009308	amine metabolic process	Biological Process	3
5	GO:0016999	antibiotic metabolic process	Biological Process	3
5	GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	Biological Process	1
5	GO:1901361	organic cyclic compound catabolic process	Biological Process	6
5	GO:1901362	organic cyclic compound biosynthetic process	Biological Process	5
5	GO:0018130	heterocycle biosynthetic process	Biological Process	5
5	GO:0060255	regulation of macromolecule metabolic process	Biological Process	5
5	GO:0043412	macromolecule modification	Biological Process	3
5	GO:0006796	phosphate-containing compound metabolic process	Biological Process	3
5	GO:0005976	polysaccharide metabolic process	Biological Process	3
5	GO:0080090	regulation of primary metabolic process	Biological Process	5
5	GO:0031323	regulation of cellular metabolic process	Biological Process	5
5	GO:0006082	organic acid metabolic process	Biological Process	3
5	GO:0019439	aromatic compound catabolic process	Biological Process	3
5	GO:0051171	regulation of nitrogen compound metabolic process	Biological Process	5
5	GO:0016052	carbohydrate catabolic process	Biological Process	3
5	GO:0046700	heterocycle catabolic process	Biological Process	6
5	GO:0006139	nucleobase-containing compound metabolic process	Biological Process	8
5	GO:0009889	regulation of biosynthetic process	Biological Process	5

Level	GO ID	Term	Type	#Seqs
5	GO:0044765	single-organism transport	Biological Process	8
5	GO:0010467	gene expression	Biological Process	5
5	GO:0019438	aromatic compound biosynthetic process	Biological Process	5
5	GO:0019538	protein metabolic process	Biological Process	11
5	GO:0043603	cellular amide metabolic process	Biological Process	3
5	GO:0009057	macromolecule catabolic process	Biological Process	6
5	GO:1901565	organonitrogen compound catabolic process	Biological Process	3
5	GO:0046677	response to antibiotic	Biological Process	3
5	GO:0009059	macromolecule biosynthetic process	Biological Process	5
5	GO:0044260	cellular macromolecule metabolic process	Biological Process	11
5	GO:0016032	viral process	Biological Process	3
5	GO:0007165	signal transduction	Biological Process	3
5	GO:0044270	cellular nitrogen compound catabolic process	Biological Process	6
5	GO:0044271	cellular nitrogen compound biosynthetic process	Biological Process	5
6	GO:0072338	cellular lactam metabolic process	Biological Process	3
6	GO:0019219	regulation of nucleobase-containing compound metabolic process	Biological Process	5
6	GO:0006811	ion transport	Biological Process	7
6	GO:0055085	transmembrane transport	Biological Process	4
6	GO:0044106	cellular amine metabolic process	Biological Process	3
6	GO:0019058	viral life cycle	Biological Process	3
6	GO:0031326	regulation of cellular biosynthetic process	Biological Process	5
6	GO:0036211	protein modification process	Biological Process	3
6	GO:0090304	nucleic acid metabolic process	Biological Process	8
6	GO:0043436	oxoacid metabolic process	Biological Process	3
6	GO:0000272	polysaccharide catabolic process	Biological Process	3
6	GO:0010468	regulation of gene expression	Biological Process	5
6	GO:0034655	nucleobase-containing compound catabolic process	Biological Process	3
6	GO:0034654	nucleobase-containing compound biosynthetic process	Biological Process	5
6	GO:0043605	cellular amide catabolic process	Biological Process	3
6	GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	Biological Process	1
6	GO:0044265	cellular macromolecule catabolic process	Biological Process	3
6	GO:0010556	regulation of macromolecule biosynthetic process	Biological Process	5
6	GO:0017001	antibiotic catabolic process	Biological Process	3
6	GO:0016311	dephosphorylation	Biological Process	3
6	GO:0034645	cellular macromolecule biosynthetic process	Biological Process	5
6	GO:0044267	cellular protein metabolic process	Biological Process	3

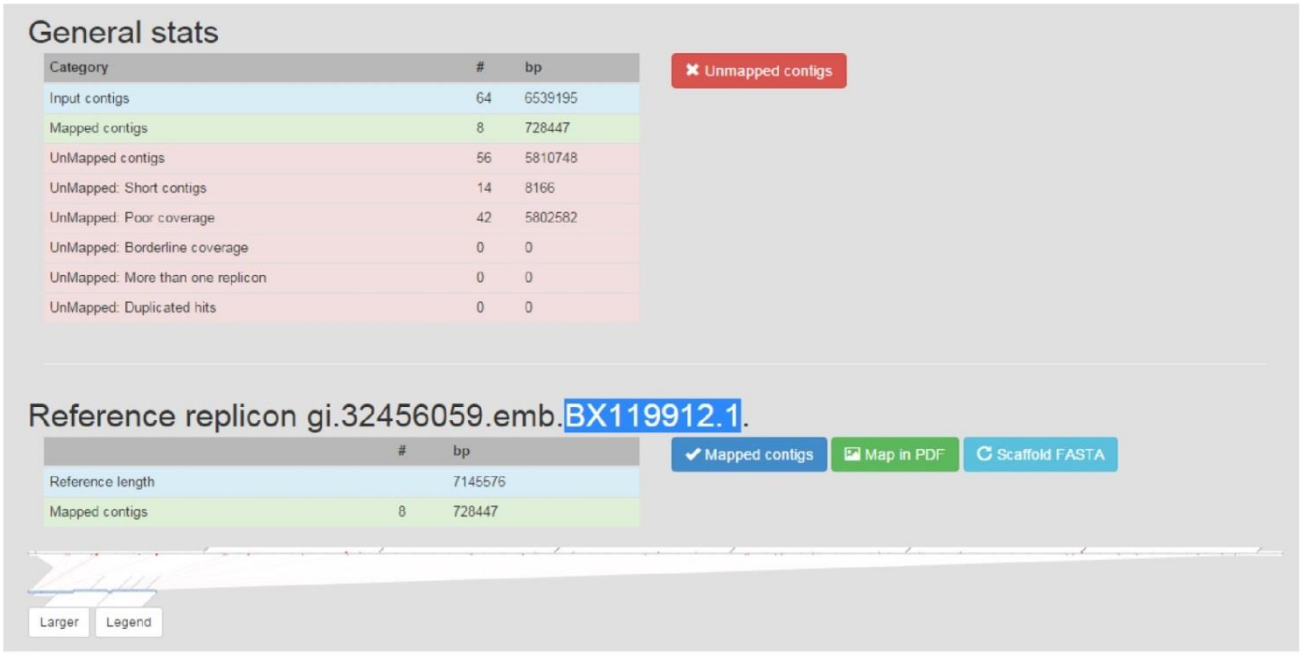
Level	GO ID	Term	Type	#Seqs
6	GO:0006508	proteolysis	Biological Process	8
7	GO:0072340	cellular lactam catabolic process	Biological Process	3
7	GO:2000112	regulation of cellular macromolecule biosynthetic process	Biological Process	5
7	GO:0019075	virus maturation	Biological Process	3
7	GO:0019752	carboxylic acid metabolic process	Biological Process	3
7	GO:0030653	beta-lactam antibiotic metabolic process	Biological Process	3
7	GO:0016070	RNA metabolic process	Biological Process	5
7	GO:0006812	cation transport	Biological Process	4
7	GO:0019068	virion assembly	Biological Process	3
7	GO:0006259	DNA metabolic process	Biological Process	3
7	GO:0006464	cellular protein modification process	Biological Process	3
8	GO:0051252	regulation of RNA metabolic process	Biological Process	5
8	GO:0030655	beta-lactam antibiotic catabolic process	Biological Process	3
8	GO:0019069	viral capsid assembly	Biological Process	3
8	GO:0006308	DNA catabolic process	Biological Process	3
8	GO:0032774	RNA biosynthetic process	Biological Process	5
8	GO:0006520	cellular amino acid metabolic process	Biological Process	3
8	GO:0006470	protein dephosphorylation	Biological Process	3
9	GO:0097659	nucleic acid-templated transcription	Biological Process	5
9	GO:2001141	regulation of RNA biosynthetic process	Biological Process	5
9	GO:0046797	viral procapsid maturation	Biological Process	3
10	GO:1903506	regulation of nucleic acid-templated transcription	Biological Process	5
10	GO:0006351	transcription, DNA-templated	Biological Process	5
11	GO:0006355	regulation of transcription, DNA-templated	Biological Process	5
11	GO:0006352	DNA-templated transcription, initiation	Biological Process	5

Appendix III. Contigs realignment results obtained from CONTIGuator mapped against *Rhodopirellula baltica* SH1^{T} genome.

LF1



UC8


General stats
Category # bp
Input contigs 64 6539195
Mapped contigs 8 728447
UnMapped contigs 56 5810748
UnMapped: Short contigs 14 8166
UnMapped: Poor coverage 42 5802582
UnMapped: Borderline coverage 0 0
UnMapped: More than one replicon 0 0
UnMapped: Duplicated hits 0 0
Unmapped contigs
Reference replicon gi.32456059.emb.BX119912.1.
bp
Reference length 7145576
Mapped contigs 8 728447
Mapped contigs
Map in PDF
Scaffold FASTA
Larger
Legend

FC18

Appendix IV. Shared proteins and subsytems shared by LF1, UC8 and FC18, retrived by RAST

Category	Subcategory	Subsystem	Role
Amino Acids and Derivatives	Alanine, serine, and glycine	Alanine biosynthesis	Cysteine desulfurase (EC 2.8.1.7), SufS subfamily
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Arginine and Ornithine Degradation	Succinylarginine dihydrolase (EC 3.5.3.23)
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Arginine and Ornithine Degradation	Succinylglutamic semialdehyde dehydrogenase (EC 1.2.1.71)
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Arginine and Ornithine Degradation	Arginine N-succinyltransferase (EC 2.3.1.109)
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Arginine and Ornithine Degradation	Arginine N-succinyltransferase (EC 2.3.1.109)
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Cyanophycin Metabolism	Cyanophycinase (EC 3.4.15.6)
Amino Acids and Derivatives	Aromatic amino acids and derivatives	Chorismate: Intermediate for synthesis of Tryptophan, PAPA antibiotics, PABA, 3-hydroxyanthranilate and more.	Isochorismatase (EC 3.3.2.1)
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Arginine and Ornithine Degradation	Succinylarginine dihydrolase (EC 3.5.3.23)
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Arginine and Ornithine Degradation	Succinylglutamic semialdehyde dehydrogenase (EC 1.2.1.71)
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Arginine and Ornithine Degradation	Succinylglutamic semialdehyde dehydrogenase (EC 1.2.1.71)
Amino Acids and Derivatives	Fermentation	Acetyl-CoA fermentation to Butyrate	Acetyl-CoA acetyltransferase (EC 2.3.1.9)
Amino Acids and Derivatives	Central carbohydrate metabolism	Pyruvate metabolism I: anaplerotic reactions, PEP	Phosphoenolpyruvate carboxykinase [ATP] (EC 4.1.1.49)
Amino Acids and Derivatives	no subcategory	Peptidoglycan Biosynthesis	Glucosamine-1-phosphate N-acetyltransferase (EC 2.3.1.157)
Amino Acids and Derivatives	Capsular and extracellular polysacchrides	Exopolysaccharide Biosynthesis	Glycosyl transferase, group 2 family protein
Amino Acids and Derivatives	no subcategory	Peptidoglycan Biosynthesis	UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase (EC 6.3.2.13)
Amino Acids and Derivatives	no subcategory	CBSS-266117.6.peg.1260	16S rRNA (guanine(966)-N(2))-methyltransferase (EC 2.1.1.171)
Amino Acids and Derivatives	no subcategory	CBSS-296591.1.peg.2330	Lipid carrier : UDP-N-acetylgalactosaminyltransferase (EC 2.4.1.-)
Amino Acids and Derivatives	Biotin	Biotin biosynthesis	3-ketoacyl-CoA thiolase (EC 2.3.1.16)
Amino Acids and Derivatives	Folate and pterines	Folate Biosynthesis	Dihydrofolate synthase (EC 6.3.2.12)
Amino Acids and Derivatives	Folate and pterines	Folate Biosynthesis	Folylpolyglutamate synthase (EC 6.3.2.17)
Amino Acids and Derivatives	NAD and NADP	NAD and NADP cofactor biosynthesis global	Nudix-related transcriptional regulator NrtR
Amino Acids and Derivatives	no subcategory	DNA structural proteins, bacterial	Integration host factor beta subunit
Amino Acids and Derivatives	no subcategory	Ton and Tol transport systems	TPR domain protein, putative component of TonB system

Category	Subcategory	Subsystem	Role
Amino Acids and Derivatives	no subcategory	Broadly distributed proteins not in subsystems	YbbL ABC transporter ATP-binding protein
Amino Acids and Derivatives	no subcategory	Broadly distributed proteins not in subsystems	YbbM seven transmembrane helix protein
Amino Acids and Derivatives	no subcategory	Broadly distributed proteins not in subsystems	YbbL ABC transporter ATP-binding protein
Amino Acids and Derivatives	no subcategory	Broadly distributed proteins not in subsystems	YbbM seven transmembrane helix protein
Amino Acids and Derivatives	Flagellar motility in Prokaryota	Flagellar motility	Signal transduction histidine kinase CheA (EC 2.7.3.-)
Amino Acids and Derivatives	Phages, Prophages	Phage capsid proteins	Phage major capsid protein
Amino Acids and Derivatives	no subcategory	Phosphate metabolism	Pyrophosphate-energized proton pump (EC 3.6.1.1)
Amino Acids and Derivatives	no subcategory	Glutathione-regulated potassium-efflux system and associated functions	Glutathione-regulated potassium-efflux system protein KefB
Amino Acids and Derivatives	Protein processing and modification	Protein deglycation	Ribulosamine/erythrulosamine 3-kinase potentially involved in protein deglycation
Amino Acids and Derivatives	Transcription	Transcription initiation, bacterial sigma factors	RNA polymerase sigma-70 factor
Amino Acids and Derivatives	no subcategory	SigmaB stress responce regulation	Anti-sigma B factor antagonist RsbV
Amino Acids and Derivatives	Oxidative stress	Glutathione: Non-redox reactions	Lactoylglutathione lyase (EC 4.4.1.5)
Amino Acids and Derivatives	Resistance to antibiotics and toxic compounds	Copper homeostasis: copper tolerance	Copper homeostasis protein CutE
Amino Acids and Derivatives	Invasion and intracellular resistance	Listeria surface proteins: Internalin-like proteins	internalin, putative